DISEÑO, VALIDACIÓN E IMPLEMENTACIÓN DE UNA HERRAMIENTA PARA LA IDENTIFICACIÓN DE METABOLITOS

Alberto Gil de la Fuente

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA. FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTESNE DE MADRID

Trabajo Fin Máster en Ingeniería Informática

Fecha

20 de junio de 2016

Directores:

Adrián Riesco Rodríguez
Abraham Otero Quintana

Autorización de difusión

Alberto Gil de la Fuente

Fecha

El/la abajo firmante, matriculado/a en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "Diseño, validación e implementación de una herramienta para la identificación de metabolitos", realizado durante el curso académico 2015-2016 bajo la dirección de Adrián Riesco Rodríguez y con la colaboración externa de dirección de Abraham Otero Quintana en el Departamento de Sistemas Informáticos y Computación, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Resumen

La Metabolómica es una sub-área de la biología de sistemas que tiene como objetivo el estudio de las moléculas de pequeño tamaño (normalmente <1000 Da) llamadas metabolitos. Los metabolitos son el resultado de las reacciones químicas que concurren en una célula y que revelan información acerca del estado del organismo en el que se encuentran. La parte computacional de un análisis metabolómico comienza con la identificación de los compuestos químicos (metabolitos) correspondientes con las masas obtenidas mediante espectrografía de masas, y se lleva a cabo mediante búsquedas manuales en múltiples bases de datos de metabolitos. El proceso de identificación requiere del análisis de cada una de las masas detectadas en el espectrómetro junto a datos que ofrece la espectrometría, como es la abundancia de cada una de las masas o los tiempos de retención. Este proceso es tedioso y consume una gran cantidad de tiempo del químico analítico, pues debe buscarse la información base de datos a base de datos e ir cruzando los datos de cada una de las búsquedas hasta obtener una lista de resultados formada por los metabolitos presentes en la muestra a analizar. El objetivo de este proyecto es desarrollar una herramienta web que simplifique y automatice la búsqueda e identificación de metabolitos. Para ello se ha construido una herramienta capaz de integrar y buscar automáticamente información de los metabolitos en múltiples bases de datos metabolómicas. Esto ha requerido unificar los compuestos entre las diferentes bases de datos cuando había suficiente información para asegurar que los compuestos provenientes de varias fuentes de datos eran realmente el mismo. Además, en este proceso de búsqueda se tiene en cuenta conocimiento sobre las reacciones químicas que pueden alterar la masa del metabolito registrada por el espectrómetro de masas, como la formación de aductos y multímeros.

Palabras clave

Metabolómica
Bases de Datos
Servidor de aplicaciones
Integración de Bases de Datos
Identificación de metabolitos
Aplicaciones web

Abstract

Metabolomics is a sub-field of systems biology which has as a target the study of molecules of small size (usually <1000 Da) named metabolites. Metabolites are the outcome of the chemical reactions that occur in a cell. They provide information about the state of the organism to which the cell belongs to. The computational analysis in a metabolomics study starts with the identification of the chemical compounds (metabolites) corresponding with the experimental masses obtained by mass spectrometry. This task is carried out through manual searches in multiple metabolomomic databases.

The identification process requires the analysis of the experimental masses detected by the spectrometer and other data such as the abundance of each mass or the retention time. This is a long and tedious process which requires a large amount of time of the analytical chemists because they have to query manually multiple databases and integrate the results from each search. The result of this task is a list that contains the metabolites that have been successfully identified. The goal of this project is to develop a web tool that simplifies and automates the search and identification of metabolites. To achieve this, a web application able to integrate and query automatically data from different metabolomic databases has been built. This required to unify compounds from the integrated databases when enough information for ensuring that compounds from different databases were actually the same compound was available. Furthermore, in the search procedure, it should be taken into account expert knowledge about chemical reactions which may change the experimental mass of the metabolite detected by mass spectrometer, such adduct or multimer formation.

Keywords

Metabolomics
Databases
Application Servers
Databases integration
Metabolites identification
Web applications

Índice general

Índice de figuras

Lista de Tablas

Agradecimientos

Me gustaría agradecer todo el esfuerzo dedicado a mis directores de proyecto: Abraham Otero Quintana y Adrián Riesco Rodríguez.

Capítulo 1

Introducción

En este primer capítulo se va a realizar una introducción en el mundo de la metabolómica para comprender el ámbito de este trabajo y las necesidades por las que surge este proyecto. También se presentará el plan de trabajo y la estructura de la memoria.

1.1. Antecedentes

1.1.1. La metabolómica

La metabolómica es una sub-área de la biología de sistemas que tiene como objetivo el estudio de las moléculas de pequeño tamaño (normalmente <1.000 Da) producidas por los procesos metabólicos que concurren en una célula[23]. Un Dalton es la doceava parte de la masa de un átomo neutro de Carbono-12 (^{12}C) y 1.000 Da equivalen a $1{,}66054 \cdot 10^{-21}$ gramos. A dichas moléculas se las conoce bajo el nombre de metabolitos, y al conjunto completo de todos los metabolitos presentes en una célula se les denomina metaboloma[37]. Los metabolitos revelan información acerca del estado y el funcionamiento del organismo en el que se encuentran[23]. Por ejemplo, estudios metabolómicos han detectado las rutas metabolómicas afectadas en ratas que sufrían hipercolesterolemia, siendo esto un riesgo para el desarrollo de enfermedades más graves, y haciendo posible una temprana detección con vistas a una corrección por medio de una variación la dieta[15]. Otro ejemplo de cómo la metabolómica ha ayudado a comprender y actuar en organismos es el éxito mostrado

por la aplicación de la mitomicina C (MMC) en pacientes con cáncer de páncreas que han desarrollado resistencia a otros tratamientos[36]. En dicho estudio se demuestra como la aplicación de MMC tiene una efectividad mayor que otros agentes como la rapamicina (RM) sobre el metabolismo del tumor.

La metabolómica es la más reciente de las distintas ciencias ómicas. Las ciencias ómicas son todas las disciplinas que estudian las funciones e interacciones en un sistema biológico. Existen multitud de ciencias ómicas, siendo, hasta la aparición de la metabolómica, la genómica, la transcriptómica y la proteómica las principales[29]. La metabolómica proporciona información sobre el estado funcional de la célula u organismo mayor que el de todas estas técnicas, ya que su estudio no se centra en un compuesto químico concreto (ácido desoxirribonucleico, ácido ribonucleico o proteínas, respectivamente), sino que estudia la abundancia de una gran variedad de compuestos químicos que son el resultado final de los procesos que concurren en una célula[56]. Se estima que el metaboloma típico de una célula contiene entre 1.000 y 200.000 metabolitos diferentes, dependiendo del organismo concreto y del tipo de célula concreta[67]. Así, por ejemplo, el metaboloma sanguíneo contiene 4.549 metabolitos diferentes conocidos en la actualidad[41].

Simplificando las ciencias ómicas, se podría decir que la genómica, que estudia las moléculas del ADN, se encarga de lo que un organismo puede hacer en potencia; la transcriptómica, que estudia las moléculas de ARN mensajero, estudia lo que las células planean hacer; la proteómica, que estudia las proteínas, en particular su estructura y función, se basa en lo que está haciendo; y la metabolómica, ciencia que nos ocupa en este proyecto, es la encargada del estudio de todos los demás tipos de moléculas presentes en la célula, tal y como se puede apreciar en la figura 1.1. A pesar de ser un campo de reciente creación, la metabolómica presenta un rápido crecimiento en la actualidad y ha demostrado ser una excelente herramienta para la investigación biomédica[25].

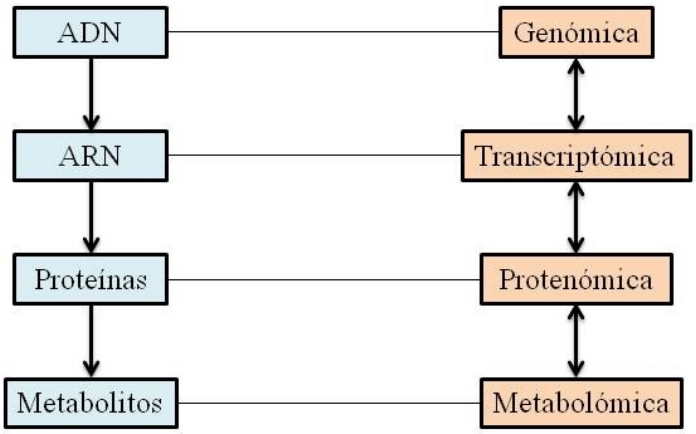


Figura 1.1: *Sub-áreas de la biología de sistemas y compuestos que estudian*

1.1.2. Análisis metabolómicos

El análisis metabolómico comienza con el proceso de preparación de la muestra. Esta preparación facilita el procesamiento de los resultados con técnicas de espectrometría de masas ya que permite aislar determinados compuestos para su análisis y se reducen los resultados generados en el espectrómetro. En consecuencia, el número de resultados a analizar es menor y se consigue un decremento del tiempo necesario para el análisis[56]. Una vez la muestra objeto del análisis ha sido extraída, aislada y estabilizada adecuadamente, existen distintas tecnologías analíticas que se pueden aplicar en metabolómica, fundamentalmente resonancia magnética nuclear (NMR) y espectrometría de masas. Mientras que la NMR permite detectar metabolitos con una masa entre 200μg y 5mg , la sensibilidad de la espectrometría de masas es mayor y es capaz de detectar metabolitos de menor peso, de 1 a 100pg (10^{-12}). Esta diferencia en la sensibilidad de la técnica (10^6 unidades) hace que la aplicación de la NMR esté limitada para metabolitos mayoritarios y la espectrometría de masas sea la teć nica utilizada para toda clase de metabolitos, siendo también utilizada para la proteómica[23].

La espectrometría de masas es una técnica analítica que consiste en la separación de

las moléculas de la muestra a analizar en función de su masa. La espectrometría de masas es utilizada en diferentes campos, pero en esta memoria se hablará de su aplicación en el ámbito de la metabolómica. La espectrometría de masas, generalmente acoplada a una etapa de separación de los metabolitos, permite eliminar aquellos compuestos químicos que no pueden ser resueltos por el detector, además de proporcionar información sobre el tiempo de retención de cada metabolito. El tiempo de retención de los metabolitos es el tiempo que tarda cada compuesto en pasar por la columna de separación que elude los compuestos en el espectrómetro. Las tres principales técnicas de separación son: cromatografía de gases (GC/MS), cromatografía de líquidos (LC/MS) y electroforesis capilar (CE/MS)[23]. En la tabla 1.1 se muestran las ventajas y limitaciones de cada una de estas opciones y el estado de las muestras que acepta cada una de ellas.

En la figura 1.2 se aprecia el funcionamiento de un espectrómetro de masas. Se introduce la muestra a analizar, se ioniza dicha muestra para la posterior separación en función de su carga, y se detectan los elementos de la muestra en función de su proporción masa/carga y

Técnica	**Ventajas**	**Limitaciones**	**Estado de la muestra**
LC/MS	Alta sensibilidad y múltiples variantes cromatográficas	Requiere gran capacidad de computación para el proceso de datos y la identificación de metabolitos es compleja	Muestras líquidas o sólidas disueltas en agua o solventes orgánicos
GC/MS	Facilidad en identificación de metabolitos y alta reproducibilidad	Solo aplicable a compuestos térmicamente estables y volátiles	Muestras gaseosas
CE/MS	El volumen de muestra necesario es pequeño	Solo aplicable a compuestos polares cargados y baja reproducibilidad	Muestras líquidas o disueltas en un capilar de separación
NMR	Altamente reproducible y sencilla preparación de la muestra	Baja sensibilidad	Muestras líquidas o sólidas

Tabla 1.1: *Ventajas e inconvenientes de diferentes métodos para estudios metabolómicos basados en espectrometría de masas*

del tiempo que ha tardado en pasar por la columna de separación. Las moléculas que tienen un peso menor o mayor al rango configurado previamente no son detectados. También existe una limitación respecto a la intensidad de los elementos, llamado umbral de detección. Hay un umbral mínimo y un umbral máximo para que los elementos sean detectados por el espectrómetro. Previamente a su introducción en el espectrómetro de masas es necesario preparar adecuadamente la muestra. La preparación de la muestra consiste en un conjunto de acciones sobre la muestra antes de analizarla en el espectrómetro de masas. Una buena preparación de la muestra facilita la obtención de resultados y el procesamiento de estos. Por ejemplo, si mediante la preparación de la muestra se logra la separación de las proteínas, moléculas que no son de interés para la mayoría de estudios metabolómicos (no así en los estudios proteómicos), el espectrómetro de masas detectará un menor número de compuestos y se facilitará el posterior análisis e interpretación de los resultados. Antes del análisis en el espectrómetro, también es necesario proceder a la ionización de los metabolitos (ver figura 1.2). La ionización es el proceso por el que las moléculas obtienen carga negativa o positiva en función de la ganancia (anión) o pérdida (catión) de electrones formando iones.

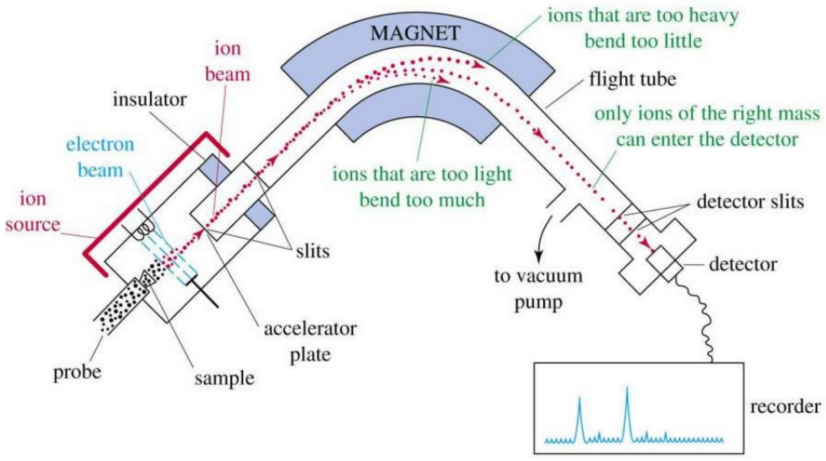


Figura 1.2: *Funcionamiento del espectrómetro de masas*[44]

Existen diferentes métodos de ionización, siendo en la espectrometría de masas los más comúnmente empleados: la ionización mediante impacto (habitualmente empleada con cro-

matografía de gases), la ionización por electro-spray (habitualmente empleada con cromatografía de líquidos) y la ionización química. La eficiencia de la ionización no es absoluta, y hay metabolitos en la muestra que pueden tener diferentes modos de ionización o pueden escapar de la misma[56]. Así mismo, en los procesos metabolómicos que se dan al procesar muestras, no todas las uniones tienen lugar, pudiendo generarse algunas muestras por debajo del umbral de detección o quedar otras por encima del umbral de saturación[56]. Esto hace que determinados metabolitos que sería esperable encontrar puedan no aparecer o que se detecten moléculas cuya cantidad exacta no pueda medirse debido a que supera la intensidad máxima detectada. De la misma forma, algunos metabolitos rompen sus enlaces durante su análisis en el espectrómetro y dan lugar a fragmentos que, idealmente, deben ser identificados como fragmentos de un metabolito precursor.

1.1.3. Análisis computacional de datos metabolómicos

Para explicar por qué se desea acelerar el tiempo que gastan los químicos intentado reducir el tiempo empleado en la etapa de la identificación se realizó un estudio a partir de la Ley de Amdahl. La Ley de Amdahl es utilizada para conocer el margen máximo de mejora de un sistema. La fórmula de la Ley de Amdahl es la siguiente:

$$T_m = T_a \cdot \left((1 - F_m) + \frac{F_m}{A_m}\right)$$

Siendo:

- T_m = Tiempo de ejecución después de la mejora.
- T_a = Tiempo de ejecución antes de la mejora.
- $F_m = \frac{Tiempo \quad del \quad subsistema \quad mejorado}{Tiempo \quad total \quad del \quad sistema}$ (Fracción de tiempo que utiliza el subsistema).

- A_m = factor de mejora del subsistema mejorado.

Generalizando la Ley de Amdahl y aplicándola a un supuesto no computacional, donde también es válida pues se trata de tareas independientes de un sistema, como máximo podría reducirse el porcentaje de tiempo con respecto al total que ocupa dicha tarea. Por tanto, para mejorar el tiempo de ejecución de un conjunto de tareas independientes secuenciales en un sistema lo más lógico es centrarse en aquella tarea que requiere una mayor proporción de tiempo.

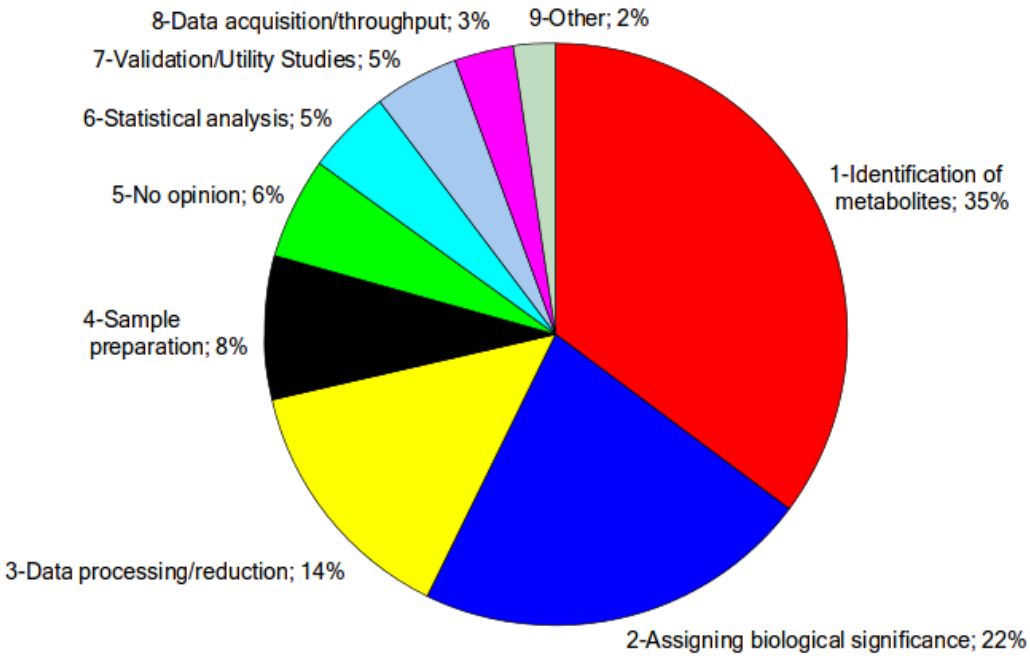


Figura 1.3: *Porcentaje de tiempo que se dedica a cada tarea en un estudio metabolómico*[33]

En este sentido, el principal cuello de botella actualmente en un estudio de metabolómica se da en la identificación de metabolitos[33]. La figura 1.3 representa la percepción de los investigadores dedicados a la metabolómica respecto a qué etapa es la que les ocupa la mayor parte del tiempo. Realizando una aproximación orientativa a partir de la figura 1.3 y aplicando la Ley de Amdahl previamente explicada, las etapas en las que se podría ob-

tener una ganancia mayor de tiempo son la identificación de metabolitos y la asignación de significado biológico. Este proyecto se va a centrar en la identificación de metabolitos, ya que, además de ser la etapa con una mayor posible ganancia potencial de tiempo, es una tarea más mecánica y que se adapta mejor a la automatización mediante herramientas computacionales que la asignación biológica.

En un análisis metabolómico, la identificación de los metabolitos comienza con la búsqueda de las masas identificadas por el espectrómetro en bases de datos de metabolitos. Sin embargo, la imprecisión de los dispositivos experimentales puede hacer que los picos del espectro de masas no coincidan perfectamente con sus valores teóricos. Además, existen una serie de efectos que alteran la masa original del metabolito. Entre estos destacan:

- Isótopos: se llama isótopos a los átomos de un mismo elemento que tienen un número diferente de neutrones, y, en consecuencia, tienen una masa atómica diferente. Un mismo metabolito puede haberse formado a partir de isótopos diferentes de un átomo, y por tanto dos moléculas de ese metabolito pueden tener masas ligeramente diferentes. Por ejemplo, la molécula de carbono (**C**) pesará 12,0107 Da si el elemento es ^{12}C, 13,0030 Da si es el isótopo ^{13}C o 14,0030 si el elemento es ^{14}C. El número que aparece delante del elemento químico (^{12}C) hace referencia al número de neutrones que tiene el núcleo del átomo correspondiente. En el ejemplo anterior el átomo de carbono podría tener 12, 13 o 14 neutrones respectivamente dependiendo del isótopo correspondiente.

- Aductos: los aductos son productos formados por la unión directa de dos moléculas. En el caso de la metabolómica, las moléculas pueden unirse con los metabolitos formando diferentes aductos. La formación de aductos es habitual durante el proceso de ionización, especialmente al emplear ionización basada en electro-spray. El aducto más común es el formado por la adición ([M+H]+) o sustracción ([M-H]-) de un protón, aunque también es habitual observar aductos basados en sodio ([M+Na]+), potasio ([M+K]+) y cloro ([M+Cl]-) debido a la ubicuidad de estos compuestos químicos en las muestras biológicas[5]. En los ejemplos aquí mostrados, una molecula llamada M se

ha unido a diferentes elementos, que son H, Na, K o Cl, quedando tras la unión con una carga positiva o negativa.

- Fragmentos: se llaman fragmentos a las moléculas que surgen a partir de la rotura de una molécula precursora. Durante el proceso de ionización, especialmente si se emplea ionización mediante impacto, es posible que se rompan algunos enlaces débiles de los metabolitos. Por tanto, en el espectro resultante además del pico correspondiente con la masa del metabolito aparecerán también picos correspondientes con sus fragmentos[4].

- Multímeros: metabolitos formados por varias moléculas del metabolito original, habitualmente dos o tres de ellas. Cuando la concentración en la muestra de un determinado metabolito es alta, es posible que se formen multímeros de dicho metabolito. Los multímeros en ocasiones presentan pérdidas neutras de masa, lo que puede hacer que su masa no sea exactamente 2x o 3x la masa del metabolito original[57].

- Diversas cargas en el ion: dependiendo de la estructura química de cada metabolito, es posible que durante el proceso de ionización se formen iones cargados con múltiples cargas (n=2, 3,...). En este caso en el espectro de masas, además del pico original del metabolito, se observarán picos situados en *m/2, m/3, ...* siendo m la masa original del metabolito.

Todos estos efectos combinados tienen como resultado que en un análisis metabolómico típico entre el 40 y el 80 % de los picos que se observan en el espectrograma de masas se correspondan con compuestos químicos derivados a partir de los metabolitos originales, lo que a menudo produce falsos positivos en la identificación de metabolitos[1,5,52]. Por otro lado, cerca del 50 % de los picos de masas considerados estadísticamente significativos en un estudio típico de metabolómica son compuestos no identificados. El fallo en la identificación puede deberse a los efectos que alteran la masa estimada del metabolito (isótopos, formación de aductos y multímeros, fragmentación y presencia de distintas cargas en el metabolito

ionizado)[48]. No obstante, con toda probabilidad, algunos de estos picos se corresponden con nuevos compuestos químicos que todavía no son conocidos.

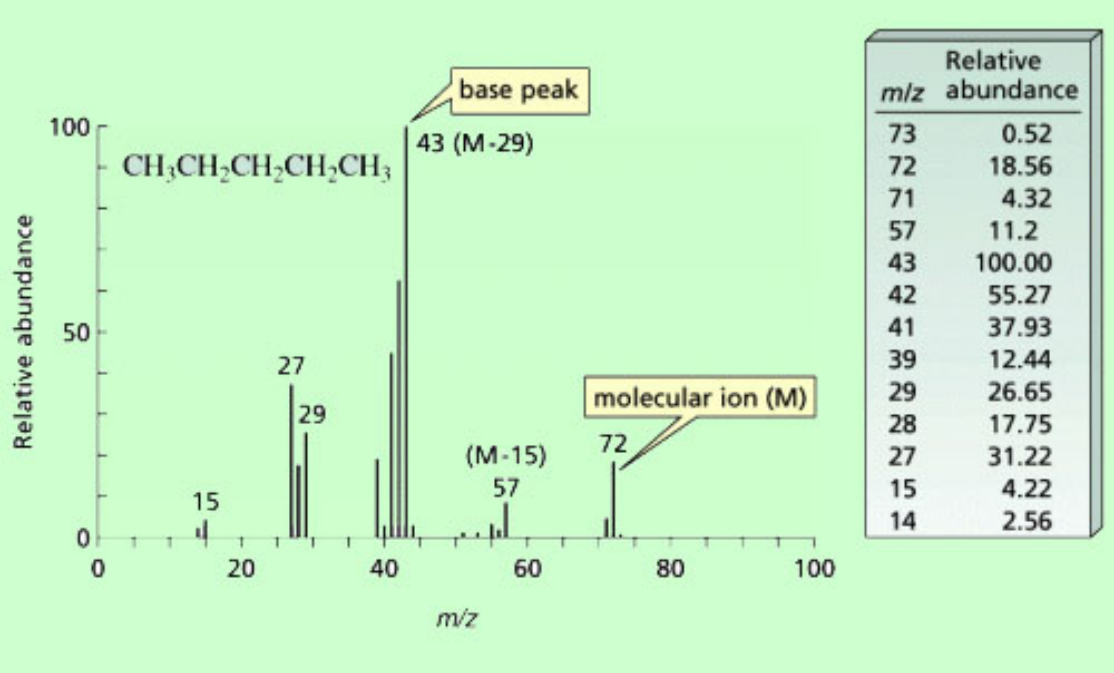


Figura 1.4: *Conjunto de picos generados en el espectrómetro de masas por un único metabolito*[12]

En la figura 1.4 se puede apreciar un ejemplo del espectro que genera un único metabolito en un espectrómetro de masas. El eje horizontal corresponde a la proporción masa/carga denominada **m/z**, y muestra el peso detectado por el espectrómetro en Daltons (**Da**). El eje vertical corresponde a la intensidad de cada uno de los pesos detectados en el eje horizontal. Es decir, detecta una abundancia de 0,52 con masa atómica 73, de 18,56 con masa atómica 72, y así sucesivamente según la tabla incluida en el gráfico. En el caso concreto de esta muestra hay un pico en torno a 72 Da identificado como el elemento con fórmula $CH_3CH_2CH_2CH_2CH_3$ del que puede deducirse que el elemento tiene tres isótopos diferentes. El más común, cuya masa es de 72, tiene una abundancia del 79,31 %, pero también se obtiene una abundancia de un 18.6 % del isótopo con masa 73 y un 0.02 % del isótopo con peso atómico de 71. Se podría analizar de forma análoga el resto del gráfico según abundancias

y masas atómicas. Estos tres isótopos del elemento surgen de los isótopos del carbono (^{12}C, ^{13}C y ^{14}C) y de los isótopos estables del hidrógeno (^{1}H y ^{2}H) - el hidrógeno tiene hasta siete isótopos, pero sólo el protio y el deuterio correspondientes a ^{1}H y ^{2}H son estables -. El resto de picos de dicho espectro pertenecen a ese mismo compuesto cuya rotura de enlaces ha generado compuestos de menor peso. El pico generado en 57 pertenece a $CH_3CH_2CH_2CH_2$, el generado en 43 a $CH_2CH_2CH_2$, el de 29 a CH_2CH_2 y, finalmente, el pico con m/z 15 corresponde a CH_2.

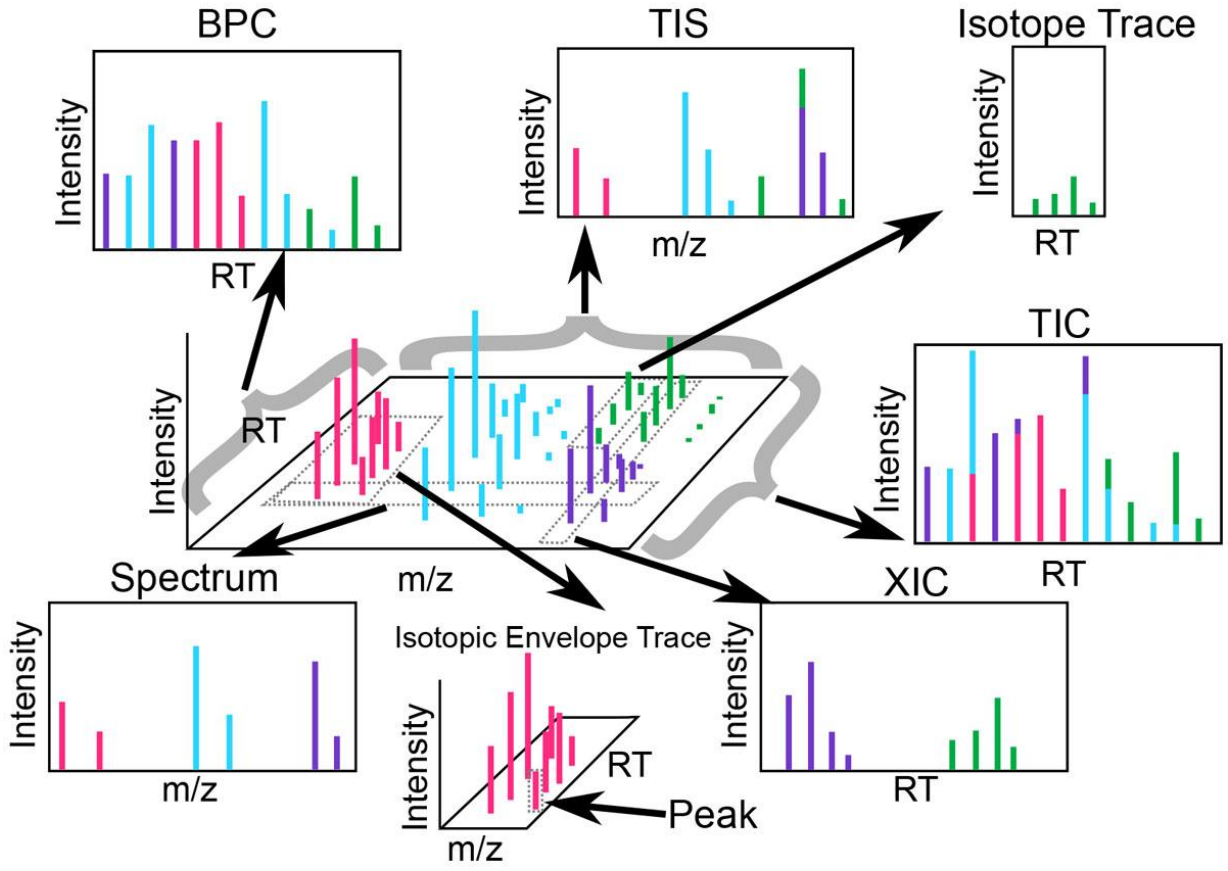


Figura 1.5: *Ejemplo de resultado generado por un espectrómetro de datos sobre una muestra en tres dimensiones (incluye tiempo de retención)*[56]

En la figura 1.5 se muestra un espectro obtenido mediante LC/MS que incluye tiempos de retención (RT). A esta técnica se la conoce como *Time-Of-Flight Mass Spectrometry (TOFMS o TOF)* o *Quadrupole-Time-Of-Flight Mass Spectrometry (QTOFMS o QTOF)*

cuando la instrumentación facilita alta precisión y alta resolución. Al considerar el RT, el espectro contiene una dimensión más que permite distinguir entre diferentes compuestos con una masa atómica igual pero diferentes tiempos de retención causados por una diferente afinidad del metabolito por la columna de separación. Esto facilita la identificación de isótopos de un mismo elemento en función del RT y de la abundancia conocida de los isótopos de cada compuesto.

Para paliar algunas deficiencias de la técnica analítica se pueden encadenar etapas del espectrómetro de masas, dando lugar a la técnica llamada espectrometría en tándem. El objetivo de esta técnica es fragmentar en sucesivas etapas los compuestos aislados en una primera etapa de espectrografía para obtener información adicional sobre su composición química a partir de estos fragmentos.

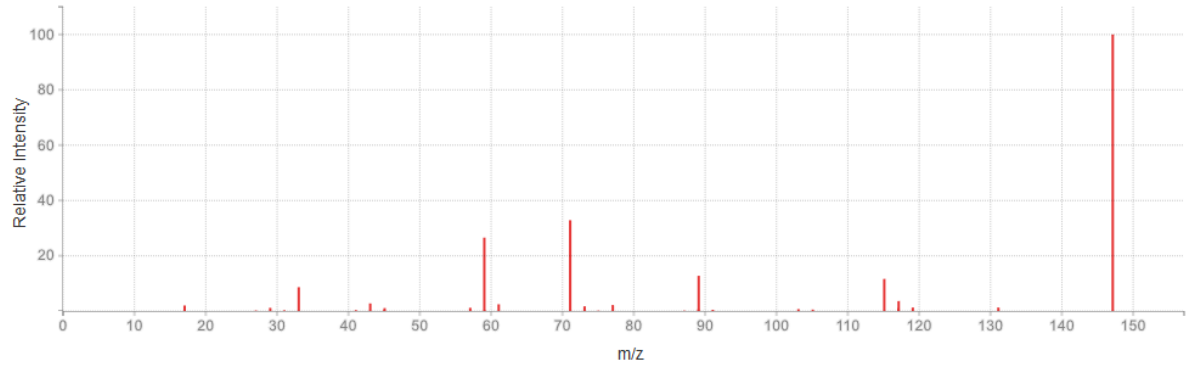


Figura 1.6: *Ejemplo de resultado de la técnica de espectrometría en tándem sobre el elemento 1-Methoxy-1-pentyloxyethane, CAS:73142-32-2*[41]

En la figura 1.6 se puede observar el espectro esperado al aplicar espectrometría en tándem con una energía de 10 voltios y modo de ionización positivo sobre el elemento *1-Methoxy-1-pentyloxyethane, CAS:73142-32-2* cuya fórmula es $C_8H_{18}O_2$ y el peso molecular es 146,1306. La molécula inicial tiene la estructura mostrada en la figura 1.7 y se descompone en estructuras más pequeñas según la rotura de sus enlaces. Dichas estructuras en las que se descompone el compuesto de origen pueden verse en el apéndice A.

Existen dos tipos de experimentos metabolómicos: los experimentos dirigidos, que bus-

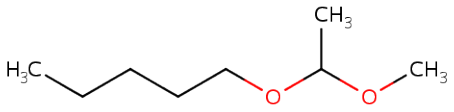


Figura 1.7: *Molécula inicial (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2) sobre la que se aplica espectrometría en tándem*[41]

can la medición de determinados compuestos previamente definidos, y los experimentos no dirigidos, que tienen como objetivo medir y comparar entre todos los metabolitos existentes en las muestras del experimento. Este proyecto está dirigido a la creación de una herramienta para la identificación de metabolitos en estudios metabolómicos no dirigidos.

En el campo médico el objetivo final de estos estudios metabolómicos no dirigidos suele ser la detección de marcadores (diferencias en metabolitos o en sus concentraciones) que difieran entre pacientes con una determinada anomalía para anticipar el tiempo de detección de dicha anomalía y tratarla con la mayor premura posible. Para ello diferentes modelos estadísticos son aplicados sobre los resultados en la búsqueda de estos marcadores de enfermedades o anomalías. Dichos modelos varían en función del estudio y están en fase de investigación según el tipo de anomalía buscada. Actualmente existen varios softwares como MzMATCH[9], MetAlign[26] o XCMS Online[61] que ayudan a la detección de marcadores a partir de muestras iniciales o de resultados ya filtrados previamente mediante métodos manuales.

En las figuras 1.8 y 1.9 se puede apreciar cómo se han encontrado dos biomarcadores diferentes en un análisis de muestras metabolómicas. El estudio consta de tres grupos: gru-

po de control (circulos blancos), grupo experimental (circulos negros) y grupo de control de calidad (triángulos). El grupo de control de calidad se utiliza para comprobar que no ha habido variación a lo largo del análisis de las muestras. Los biomarcadores o marcadores biológicos se refieren a evidencias médicas que pueden ser medidas y son reproducibles[60]. Estas evidencias médicas son cualquier sustancia, estructura o proceso detectado en el organismo que tiene influencia o predice una consecuencia o enfermedad[60]. La identificación de biomarcadores en metabolómica se refiere metabolitos presentes en el grupo de control que no se encuentran el grupo experimental o viceversa, o una diferencia considerable de concentración en alguno de ellos. Cuando ésto ocurre, dicho metabolito es considerado un biomarcador. Cada m análisis se analiza una muestra del grupo de control. En caso de que alguna muestra del grupo de control tuviera una desviación excesiva, lo habitual es descartar los análisis entre las muestras de control con desviación o, si son demasiadas, repetir el experimento al completo. A partir de la identificación de compuestos se han aplicado diferentes modelos estadísticos buscando diferencias significativas entre el grupo de control y el grupo experimental. Los ejes X e Y no tienen una información definida, sino que son reducciones de N a 2 dimensiones mediante análisis de componentes principales. Se han encontrando dos biomarcadores, uno separado por el eje X y otro separado por el eje Y.

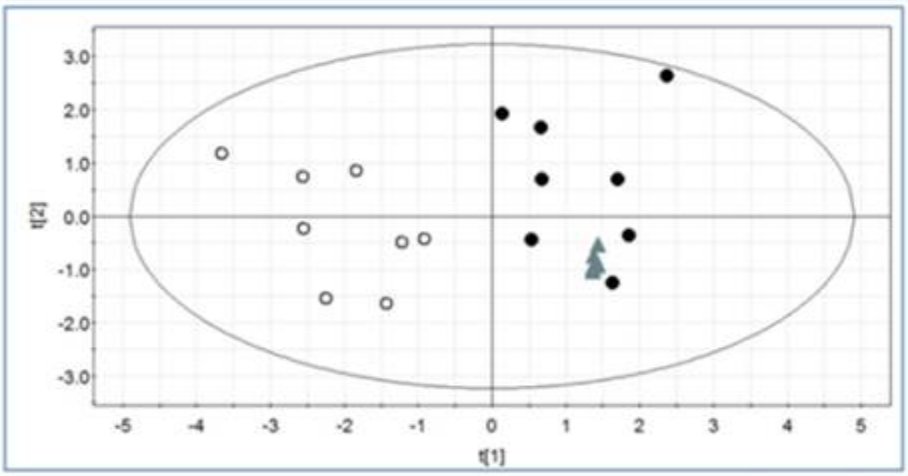


Figura 1.8: *Biomarcadores detectados tras análisis metabolómico en un ensayo clínico(1)*

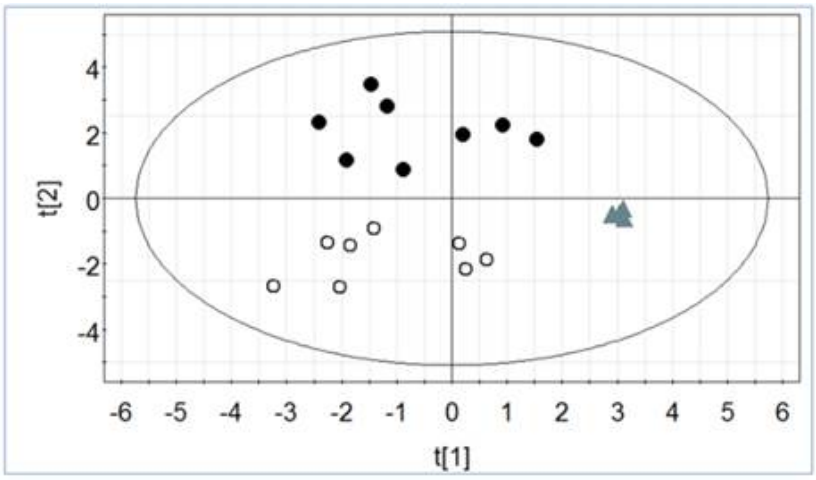


Figura 1.9: *Biomarcadores detectados tras análisis metabolómico en un ensayo clínico(2)*

1.2. Desarrollo existente

Previamente a la realización de este proyecto, existía una herramienta básica que hacía búsquedas de metabolitos en tres bases de datos (KEGG, Metlin y LipidMaps). Esta herramienta fue desarrollada por el grupo de Ingeniería Biomédica de la Universidad San Pablo-CEU en el año 2013, concretamente por el profesor Mariano Fernández López, y su misión era realizar una única búsqueda para obtener resultados de diferentes bases de datos de metabolitos. La herramienta desarrollada en este proyecto partió de este desarrollo existente. Existen otras herramientas disponibles para la identificación de metabolitos a partir de la espectrometría de masas que se explican en la sección 2.3. La herramienta recopilaba datos de dichas fuentes y los mostraba en un mismo conjunto de resultados, todos ellos tratados como entes independientes. Es decir, si un mismo compuesto aparece en las tres bases de datos, la herramienta generaba tres resultados de búsqueda diferentes que tenían que ser filtrados de forma manual posteriormente. El front-end estaba desarrollado en JavaEE y mezclaba elementos de JSF y JSP con páginas diseñadas mediante XHTML y HTML plano, sin validación de elementos ni en cliente ni en servidor. El back-end estaba desarrollado en Java. Consistía en una serie de paquetes para descargar la información de las fuentes median-

te las APIs que facilitan (KEGG y Metlin) o ficheros con la información (LipidMaps). Una vez descargada la información empleaba un analizador sintáctico para incluirla en una base de datos. El modelo entidad-relación que utilizaba se puede ver en la figura 1.10. Cada vez que había que actualizar la información de las bases de datos origen había que descargarse manualmente los recursos y ejecutar el proyecto de Java de actualización de base de datos para incluir la nueva información. La copia de seguridad de los datos se hacía manualmente.

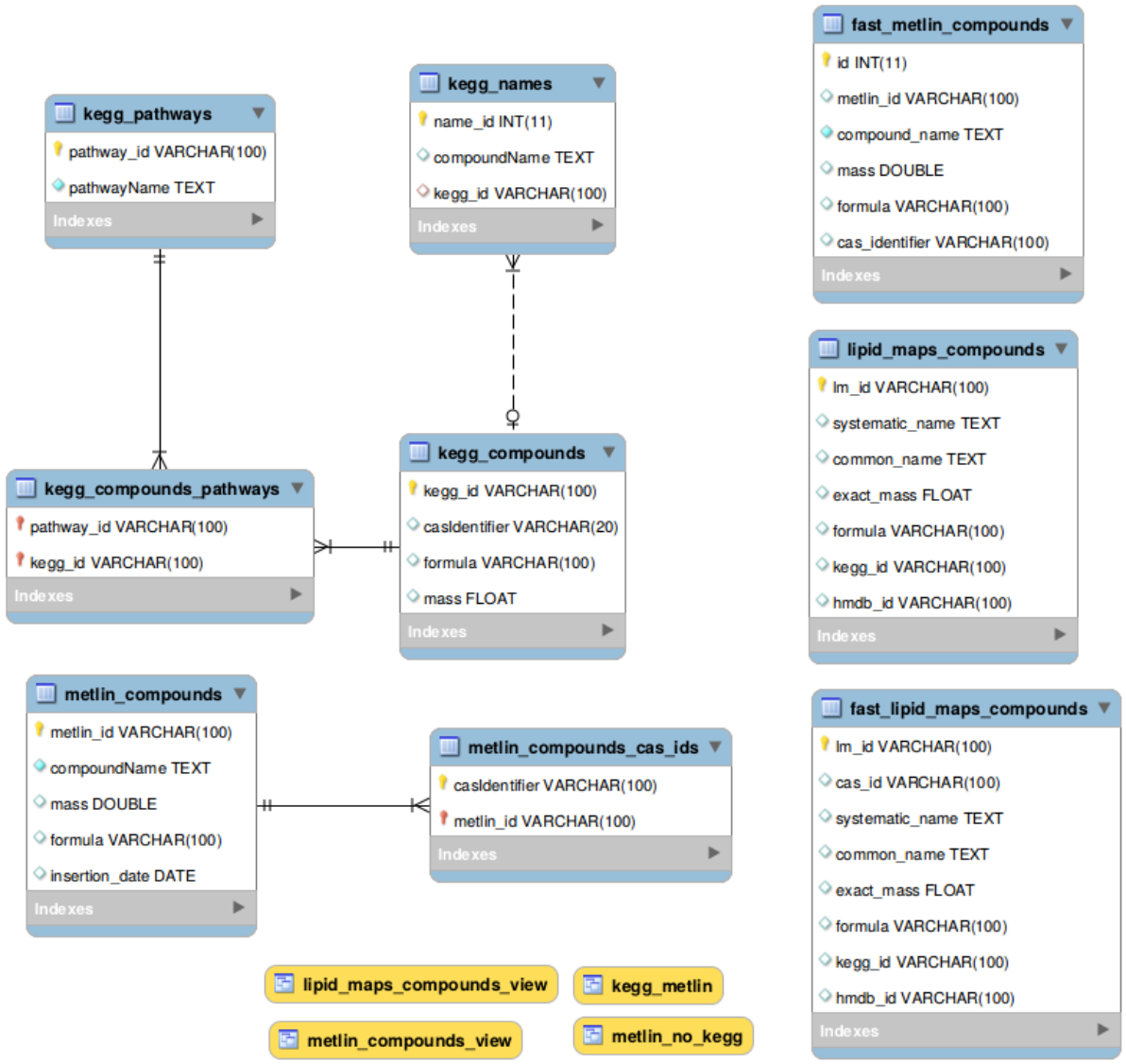


Figura 1.10: *Modelo entidad-relación de la aplicación previa*

Tras el análisis del código se decidió que para realizar la unificación de compuestos era necesario implementar un nuevo modelo de datos, así como modificar la lógica de negocio. Además de ello debían unificarse las tecnologías utilizadas en el front-end para tener mayor control y poder explotar los beneficios de CSS3 y el uso de plantillas y componentes reutilizables. Anteriormente se utilizaba una hoja de estilos, pero realmente la mayor parte de estilos estaban aplicados en línea, no haciendo uso de dicha hoja de estilos, sino en el atributo de estilo ("style") del componente. También era deseable una unificación de tecnología para la interfaz web, así como incluir validaciones tanto en cliente como en servidor para un mejor funcionamiento de la aplicación y para poder aportar información al usuario sobre posibles errores cometidos.

En cuanto al back-end, además de la unificación de compuestos, el equipo investigador tenía un gran interés en incorporar la base de datos HMDB [41] y referencias de los compuestos a la base de datos de Pub Chemical[39] (si bien no era un objetivo integrar esta última base de datos por la cantidad de información que tiene, en su mayor parte sin relación con la metabolómica). Se quería dar la posibilidad de descargar el resultado de una búsqueda en ficheros con formato .xls y además poder realizar un análisis de rutas metabólicas (*pathways*). Las rutas metabólicas son grafos de reacciones químicas definidas en diferentes mapas. En la figura 1.11 puede verse la ruta metabolómica referente a la glicólisis, obtenida de KEGG. La ruta comienza con un determinado compuesto y, en función de las reacciones químicas, el compuesto se ve alterado de forma temporal o final, en caso de que ya no sufra más transformaciones. Estas rutas pueden estar relacionadas entre sí en mapas más grandes que muestran relaciones entre las rutas concretas.

El análisis que se desea en este proyecto consiste en una ordenación de compuestos encontrados en la búsqueda de acuerdo a las posibles rutas donde está categorizado como presente. Esta funcionalidad constituye un primer paso hacia la interpretación biológica de los resultados del análisis metabolómico. Los usuarios también querían funcionalidades de búsquedas más avanzadas, como poder hacer una búsqueda incluyendo sólo determinadas

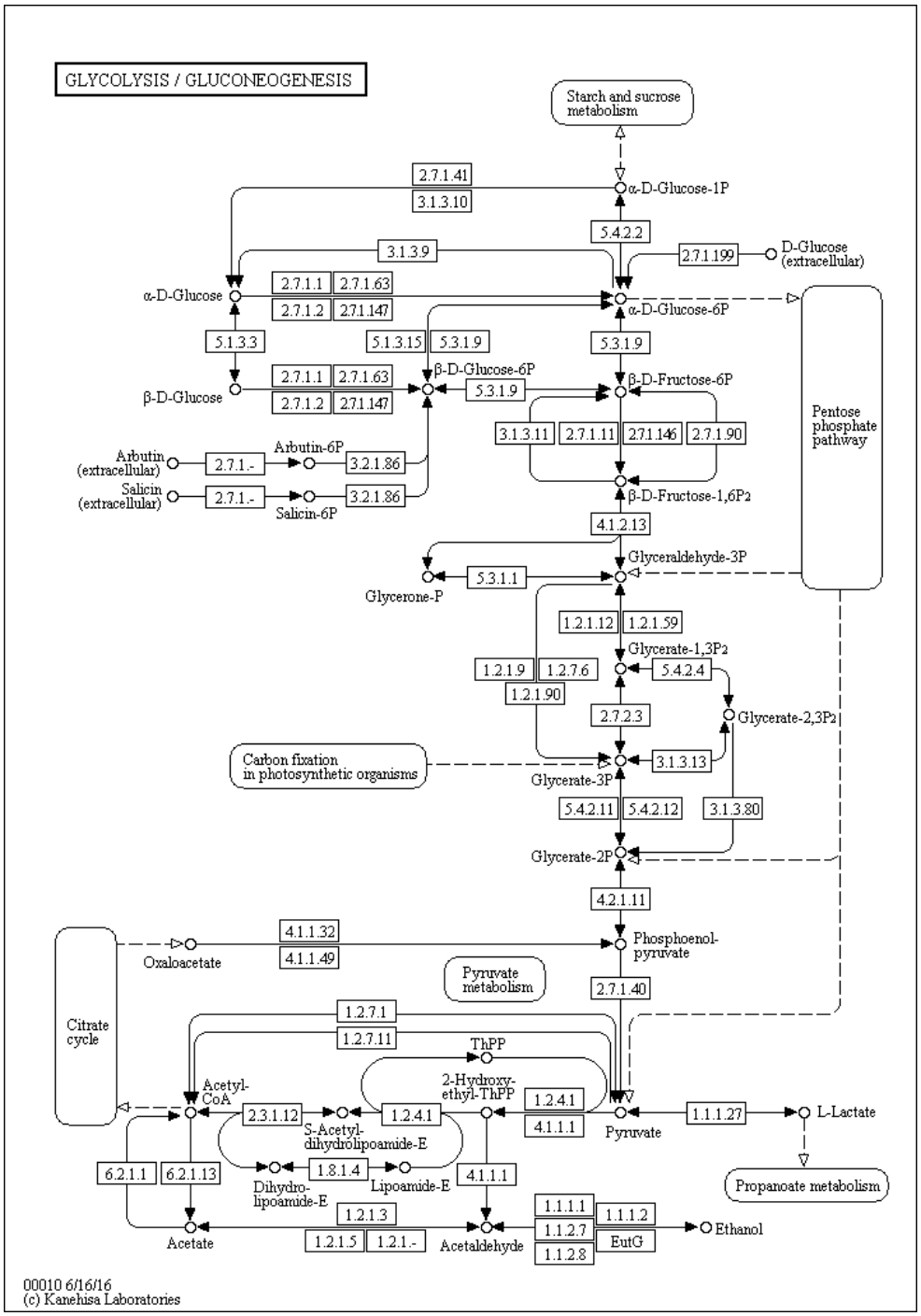


Figura 1.11: *Ejemplo de ruta metabolómica referente a la glicólisis*[40]

bases de datos y poder filtrar los elementos que forman los compuestos. En ocasiones el químico analítico sólo busca metabolitos que siguen la regla denominada CHNOPS (Carbon,

Hydrogen, Nitrogen, Oxygen, Phosporus, Sulfur), o CHNOPS y Cloro, elemento comúnmente utilizado para la ionización por electro-spray. También era deseable poder realizar búsquedas a partir de masas neutras, de compuestos con determinada carga (m/z) y también de compuestos m/z cuya masa ha sido recalculada directamente por el software de adquisición de datos del espectrómetro. Para finalizar, era deseable que las actualizaciones de datos de las bases de datos integradas estuviesen automatizadas y el sistema de copias de seguridad también, para evitar posibles errores humanos.

1.3. Objetivos

En un análisis metabolómico, la identificación de los metabolitos correspondientes con los picos obtenidos en el espectrograma de masas y la subsecuente interpretación biológica de los resultados, son tareas que requieren una gran cantidad de trabajo manual y para las cuales no existen protocolos bien definidos[13,19,51]. Estas tareas consumen una gran cantidad de tiempo del químico analítico, y por tanto de dinero, siendo actualmente los principales cuellos de botella en los análisis metabolómicos[33]. Por ello es necesario el desarrollo de estrategias tanto *in silico* como experimentales que permitan realizar búsquedas automatizadas en bases de metabolitos ya conocidos, teniendo en cuenta sus isótopos, aductos, multímeros y la posible fragmentación del metabolito. Dichas estrategias deberán estar soportadas por herramientas software que permitan automatizar y acelerar los análisis.

El objetivo de este proyecto está enmarcado como una primera parte de una tesis doctoral consistente en desarrollar estrategias y algoritmos que den soporte a la identificación de metabolitos en análisis metabolómicos, y desarrollar herramientas software que den soporte a dichas estrategias. En concreto, el trabajo va a centrarse en construir una herramienta web para reducir el tiempo que emplean los químicos en estudios metabolómicos en obtener la identificación de compuestos a partir de los resultados obtenidos en el espectrómetro de masas. Para ello, se plantean los siguientes objetivos:

1. Integración de las siguientes bases de datos de metabolitos:
    - KEGG.
    - Metlin.
    - LipidMaps.
    - HMDB.
2. Unificación en un único modelo de datos de compuestos procedentes de dichas bases de datos.
3. Renovación de la interfaz web de la herramienta utilizada hasta el momento.
4. Incorporación de conocimiento experto procedente de los químicos analíticos para filtrar los resultados de la búsqueda.
5. Exportación de resultados generados a ficheros .xls, al ser el método habitual de trabajo de los químicos.
6. Agrupación de los metabolitos identificados en función de las rutas metabólicas en las que están presentes los metabolitos encontrados durante la búsqueda.
7. Filtración de resultados en función del origen de los metabolitos (es decir, decidir en qué bases de datos específicas se va a realizar la búsqueda).
8. Búsquedas automáticas en función de los posibles aductos formados para modos de ionización positivos y negativos.
9. Filtrado de los resultados según los elementos químicos que lo forman.

1.4. Motivación

El presente Trabajo de Fin de Máster en Ingeniería Informática ha surgido por el interés del Centro de Excelencia Metabolómica y Bioanálisis de la Universidad San Pablo CEU

(CEMBIO, http://www.metabolomica.uspceu.es/) en disponer de una herramienta que proporcione soporte la identificación de metabolitos, ahorrando tiempo del químico analítico.

El proyecto tiene como finalidad desarrollar una herramienta web que permita disminuir el tiempo dedicado por los investigadores de metabolómica a la identificación de los resultados obtenidos en el análisis de metabolitos mediante espectrometría de masas. Según datos aportados por el CEMBIO, el filtrado de resultados para el estudio de una única muestra puede suponer actualmente entre uno y dos meses de trabajo de un investigador, en función del número de resultados extraídos a partir de la muestra y la experiencia del investigador. Este proceso es bastante mecánico y consiste en buscar en las diferentes bases de datos posibles metabolitos que concuerden con la información del espectrómetro e información propia de cada experimento. Esta última afirmación se refiere a que los metabolitos tienen determinadas características intrínsecas, como puede ser el origen donde puede encontrarse (o, al menos, donde ha sido identificado hasta el momento).

Una herramienta capaz de integrar información sobre metabolitos extraída de múltiples bases de datos de metabolómica y con capacidad para buscar sobre los datos integrados no sólo las masas originales, sino posibles transformaciones (isótopos, aductos, etc.) de las masas originales y aplicar después filtros sobre ellas ahorraría una gran cantidad de trabajo manual al químico analítico.

1.5. Plan de trabajo

El plan de trabajo propuesto para este proyecto se estructura en las siguientes actividades:

1. Análisis bibliográfico relativo a la metabolómica y la identificación de metabolitos. Contacto con el campo sobre el que se va a desarrollar la herramienta para poder identificar características y necesidades del proyecto. Evaluación de herramientas disponibles en la biografía para la identificación de metabolitos.

2. Diseño de la herramienta. Diseño de la lógica de negocio y creación de prototipos (*Wireframing*) de la interfaz web.

3. Implementación de back-end:

    - Elección de tecnologías.
    - Implementación de nuevo modelo de datos y de lógica de negocio.
    - Integración de bases de datos de metabolitos para la posterior identificación.
    - Unificación de compuestos extraídos de las diferentes bases de datos.
    - Validación de formularios en el servidor.

4. Implementación de front-end:

    - Unificación de tecnología a utilizar (JSF).
    - Modelo adaptativo (*responsive*) XHTML con CSS3.
    - Mayor uso de AJAX y mejora de la interfaz en general.
    - Validación de formularios en el cliente.

5. Búsqueda de aductos a partir de las señales generadas por el espectrómetro de masas e identificación de posibles rutas metabólicas que han podido seguir los elementos que forman el compuesto.

6. Escritura de la memoria.

En el diagrama de Gantt mostrado en la figura 1.12 se observa la planificación llevada a cabo en el proyecto. Los ciclos (*sprints*) de trabajo del proyecto se dividieron por semanas, comenzando en febrero y con fecha límite el día 20 de junio, dejando una semana libre para ajustar posibles retrasos en alguna de las tareas. El primer mes estaba planificado para el análisis bibliográfico y la evaluación de herramientas disponibles en el mercado. A partir de la cuarta semana, debía comenzarse el diseño de la herramienta, realizando el correspondiente

diseño de prototipos que se mostrará en la sección 3.2. A partir de entonces comenzaría la implementación de la herramienta, cuyo desarrollo estaba planificado en dos meses. El primer mes se destinaría al desarrollo del back-end, y el segundo mes a la implementación del front-end y las nuevas funcionalidades solicitadas en el informe de requisitos. La escritura correspondía a las tres últimas semanas del proyecto, dejando la última semana sin tareas para posibles contratiempos en las tareas anteriormente mencionadas.

Actividad	Febrero	Marzo	Abril	Mayo	Junio	Julio
1. Análisis bibliográfico						
1.a. Identificación de metabolitos	[X][X] | |	| | | | |	| | | | |	| | | | |	| | | | |	| | | | |
1.b. Evaluación de herramientas	| | [X][X]	| | | | |	| | | | |	| | | | |	| | | | |	| | | | |
2. Diseño de la herramienta						
2.a. Diseño lógica de negocio	| | | | |	[X][X][X] |	| | | | |	| | | | |	| | | | |	| | | | |
2.b.Wireframe	| | | | |	| | [X][X]	| | | | |	| | | | |	| | | | |	| | | | |
3. Implementación de back-end						
3.a. Elección de tecnologías	| | | | |	| | | | |	[X] | | |	| | | | |	| | | | |	| | | | |
3.b. Implementación de nuevo modelo de datos y de lógica de negocio.	| | | | |	| | | | |	[X][X][X][X]	| | | | |	| | | | |	| | | | |
3.c. Integración de bases de datos de metabolitos para la posterior identificación.	| | | | |	| | | | |	| | | | |	[X][X] | |	| | | | |	| | | | |
3.d. Unificación de compuestos extraidos de las diferentes bases de datos para la indentificación.	| | | | |	| | | | |	| | | | |	[X][X] | |	| | | | |	| | | | |
3.e. Validación de formularios en servidor.	| | | | |	| | | | |	| | | | |	| | [X] |	| | | | |	| | | | |
4. Implementación de front-end						
4.a. Unificación de tecnología a utilizar (JSF)	| | | | |	| | | | |	| | | | |	| | [X] |	| | | | |	| | | | |
4.b.Modelo responsive XHTML con CSS3	| | | | |	| | | | |	| | | | |	| | | [X]	[X] | |	| | | | |
4.c. Utilización de AJAX.	| | | | |	| | | | |	| | | | |	| | | [X]	[X] | |	| | | | |
4.d Validación de formularios en cliente	| | | | |	| | | | |	| | | | |	| | | [X]	[X] | |	| | | | |
5.Implementación de funcionalidades						
5.a. Implementación de búsqueda de aductos.	| | | | |	| | | | |	| | | | |	| | | | |	| [X][X] |	| | | | |
5.b. Análisis de pathways.	| | | | |	| | | | |	| | | | |	| | | | |	| [X][X] |	| | | | |
6. Escritura de memoria						
6.a. Escritura	| | | | |	| | | | |	| | | | |	| | | | |	| | | [X]	[X][X] | |

Figura 1.12: *Diagrama de Gantt*

La gestión del proyecto se ha realizado bajo una metodología Scrum, un modelo de desarrollo ágil e incremental basado en *sprints*. Este era el modelo que mejor se adaptaba a las circunstancias del mismo (ver figura 1.13) ya que en esta metodología no se cuenta con una planificación extensamente detallada, sino que se cuenta con un listado de características y requisitos que el producto debe cumplir, y dichas características se abordan en los diferentes *sprints*. Define tres roles dentro del equipo:

- *Product owner*: miembro que recibe información del producto y lo formaliza en la especificación de requisitos, incluyendo orden de prioridad.
- *Scrum Master*: miembro encargado de que el *sprint* se realice correctamente.
- Equipo: responsable de desarrollar el producto cumpliendo las especificaciones.
- *Stakeholders*: participantes del proyecto. En este caso el cliente, los químicos analíticos del CEMBIO, participó de forma activa en la generación de los requisitos y en diferentes reuniones para comprobar el desarrollo de la herramienta.

En el caso de este proyecto los roles definidos por esta metodología fueron abordados por el autor y el director debido a la limitación de personas. Los dos se encargaron de acudir a las reuniones para realizar la especificación de requisitos, acordaron con el CEMBIO el orden de prioridad de las tareas y se encargaron de comprobar la correcta realización de los requisitos. El desarrollo está realizado íntegramente por el autor. En cuanto a las reuniones diarias que se recomiendan por esta metodología, se realizaban con una ventana de tiempo más holgada en función de los avances en el desarrollo. A las revisiones al terminar cada uno de los *sprints* también acudió el cliente. Los responsables del CEMBIO tenían una gran cantidad de ideas sobre diferentes automatizaciones que desean realizar, así como del procesamiento posterior de los resultados. Al no ser posible abordar todas las tareas en paralelo, la metodología Scrum fue la opción escogida para ir cumpliendo los hitos generados y continuar de forma incremental con el desarrollo del proyecto.

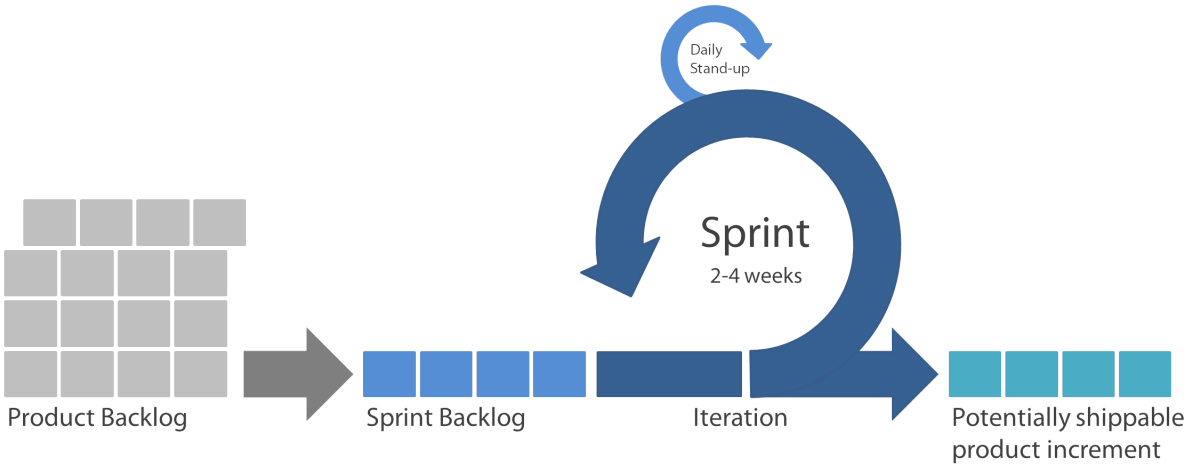


Figura 1.13: *Metodología Scrum*[20]

En las primeras reuniones del proyecto, las personas del CEMBIO tenían una hoja de ruta con varias actividades que había que ir desarrollando en función de la prioridad, por lo que se ha creado una cola de tareas en función de la prioridad con dos estados principales ("por ejecutar" y "en espera") y cinco órdenes de prioridad ("muy urgente", "urgente", "normal", "baja" y "muy baja") que permite hacer un seguimiento en función de la prioridad de las tareas y tener las mismas separadas en función de si pueden ser abordadas en el momento o están bloqueadas en espera de alguna acción que no depende del equipo de desarrollo. Una vez las tareas están finalizadas pasan al histórico. La forma de contacto con las personas involucradas ha sido el correo electrónico y las reuniones personales.

Además de las reuniones extraordinarias por cambio en la prioridad de los requisitos o según hitos conseguidos, se establecieron reuniones periódicas cada dos semanas (revisión de cada *sprint*), ajustando la fecha previamente según calendario académico y laboral de las personas involucradas. En estas reuniones periódicas se evaluaba la dirección del trabajo realizado y si había que redirigir o modificar las prioridades de las tareas. Cuando las tareas desarrolladas no se ajustaban a los requisitos, se modificaban para que cumpliese con las nuevas expectativas en caso de ser posible. Si el cambio no era posible, se buscaba alternativa para satisfacer sus necesidades.

1.6. Estructura de la memoria

El proyecto se divide en los siguientes capítulos:

I. Capítulo 1: Vista general del proyecto donde se presentan los antecedentes en la materia, la motivación, los objetivos y el plan de trabajo desarrollados.

II. Capítulo 2: Estado actual de las metodologías y herramientas para la identificación de metabolitos analizados mediante espectrometría de masas, centrándose en la técnica de cromatografía líquida.

III. Capítulo 3: Diseño de la herramienta para la identificación de metabolitos.

IV. Capítulo 4: Back-end. Capítulo dedicado al motor de la herramienta web, al mantenimiento de los servidores donde está alojada y a la lógica de negocio.

V. Capítulo 5: Lógica de presentación. Desarrollo de la interfaz web para adaptarla al prototipo desarrollado en el capítulo 3 y a la lógica de negocio implementada en el capítulo 4.

VI. Capítulo 6: Análisis de resultados obtenidos durante el proyecto y posibles líneas futuras del proyecto.

El proyecto relatado en esta memoria ha concluido con una herramienta web disponible en la dirección http://ceumass.eps.uspceu.es/mediator/index.xhtml. El código generado está disponible en dos repositorios de Bitbucket: un proyecto para la recogida de datos de las diferentes fuentes, disponible en https://bitbucket.org/albertogilf/dbupdater; y otro proyecto, accesible en https://bitbucket.org/albertogilf/ceumassmediator, que contiene el código de la aplicación web.

Capítulo 2

Estado del arte

En este capítulo se hará un breve repaso de la historia de la metabolómica y la espectrometría de masas para conocer el estado en que se encuentran ambas disciplinas (los avances en metabolómica han ido unidos, históricamente, a los avances en la instrumentación analítica). Se describirán las bases de datos y las herramientas más importantes en el ámbito de trabajo del proyecto y se hablará del estado en que se encontraba la herramienta anteriormente disponible.

2.1. Historia de la metabolómica y de la espectrometría de masas

La metabolómica surge como un avance de la bioquímica clásica cuyo objetivo es poder medir de forma cuantitativa el conjunto de metabolitos endógenos y exógenos resultantes de la actividad celular[38]. Esto es, las moléculas que están formando las diferentes células de una muestra y los cambios que se producen en función de la actividad interna y estímulos externos como pueden ser la dieta, el tabaquismo u otro tipo de acciones. El intento de detectar cambios en los organismos para el diagnóstico no se considera algo novedoso, pues ya en el siglo XVII Santorio Santorio publicó un trabajo donde trataba de obtener datos físicos a partir de la diferencia entre el peso de la comida ingerida y la expulsada y estudiar las patologías asociadas a esta diferencia[50]. En el siglo siguiente, Matthew Dobson realizó el primer análisis de orina y consiguió diferenciar los tipos de diabetes a partir del

nivel de azúcar que contenía. Estos análisis se realizaban con instrumentación adaptada de otros campos, pues no existían instrumentos diseñados de forma específica para este tipo de estudios[43].

Los estudios metabolómicos están condicionados por la exactitud de la instrumentación utilizada. Es por ello que se va a hacer un resumen de los principales hitos acontecidos con la espectrometría de masas, método mayormente utilizado en la actualidad en la metabolómica, mostrando un especial interés a la aplicación de la espectrometría al ámbito metabolómico. El primer espectrómetro de masas se creó en 1913 en la universidad de Cambridge por J.J. Thompson y fue llamado parábola de rayos positivos (*positive-ray parabola*). Nació con el objetivo de descubrir la naturaleza de los rayos catódicos. En un principio trabajaban midiendo la proporción carga-masa, medición inversa a la que suele realizarse actualmente, correspondiente a la proporción masa-carga. Este dispositivo fue el utilizado para descubrir isótopos del Neon (Ne) a partir de las masas extraídas con el espectrómetro[54]. Más adelante, en 1932, se analizó la urea de diferentes animales aplicando espectrometría de masas y se obtuvo el conocimiento del ciclo que transforma el ion amonio (NH3) en urea ((NH2)2CO). Este ciclo fue el primer ciclo metabolómico descubierto. En 1943 se describió por primera vez el antimetabolito como una sustancia que reemplaza o inhibe un metabolito específico[69]. Estos antimetabolitos tienen estructuras similares al metabolito con el que interfieren, pero reaccionan de diferente forma sobre los agentes. Un estudiante amplió el trabajo sobre el instrumento y se le atribuyó ya el nombre que conserva actualmente: espectrómetro de masas.

En la década de 1940 se utilizó el espectrómetro de masas en la industria petrolífera para medir la abundancia de pequeños hidrocarburos en los procesos de refinado. Estas mediciones eran estudios para una medición cuantitativa (estudios con objetivos concretos y dirigidos). También en 1946 se introdujo la resonancia magnética nuclear y en 1951 fueron publicados los primeros estudios sobre la relación entre los patrones metabolómicos y las enfermedades en los seres humanos[64]. En este tiempo, todas las técnicas analíticas (LC/MS,

GC/MS y NMR) estaban ya disponibles para los investigadores, y antes de 1957 se habían elucidado rutas metabólicas para las moléculas biológicas[17]. En los años 60 y 70 los químicos comenzaron a comprender la complejidad de la fragmentación de las moléculas dentro del espectrómetro[16]. Concretamente, fue Klauss Biemann en el departamento de química del MIT (**M**assachusetts **I**nstitute of **T**echnology) quién hizo uso del espectrómetro para la identificación de estructuras, estudio con el que consiguió descubrir reglas de cómo se fragmentan los péptidos y los alcaloides. La primera ocasión en que fue introducida la ionización química en la espectrometría de masas fue en 1966, cuando los científicos M.S.B. Munson y F.H. Field presentaron una publicación basada en la formación de iones obtenidos a partir de material desconocido en fase gaseosa[34]. Se dieron cuenta que el espectro que obtenían era completamente diferente del generado mediante métodos de impacto de electrones tradicionales, que era el método más utilizado en la época para conocer la estructura de los compuestos. En este mismo año se hace referencia al concepto de la metabolómica basada en espectrometría de masas y utilizada para la detección de metabolitos en una muestra de orina. Dalgliesh llevó a cabo por primera vez un experimento de GC-MS para la separación y detección de varios metabolitos presentes en tejidos biológicos y orina[7]. La primera referencia a la técnica de electro-spray data de 1968, momento en que se diluyó una solución de polímeros en una cámara de evaporación y se produjeron iones negativos y un haz de moléculas formado por dichos iones[28]. La técnica de ionización mediante electro-spray (ESI) fue desarrollada en 1989 y es la piedra sobre la que pivota la metabolómica en la actualidad. En 1994 se realizó probablemente el primer estudio basado en la técnica de cromatografía líquida mediante ionización por electro-spray. En dicho estudio se analizaron dos grupos de gatos. A los integrantes del primero de ellos se les privó de sueño, y a los del segundo se les dejó descansar de forma natural. Se descubrió un lípido desconocido que estaba presente en el grupo de gatos a los que privaron de sueño (grupo experimental) y que no estaba en el grupo de gatos sin falta de sueño (grupo de control)[21]. En la tabla 2.1 se pueden ver los principales avances en la instrumentación de la espectrometría de masas en el siglo XX.

Avance	Autor	Año
Primer método para electro-ionización mediante impacto aplicado aplicando sólidos	Dempster	1918
Primer método para electro-ionización mediante impacto aplicado aplicando gases	Bleakney	1929
Primera analizador TOF	Stephens	1946
primer analizador QTOF	Paul	1953
Ionización química	Field	1966
Primera ionización mediante desorción de campos (Aplicación de campos eléctricos sobre las muestras a analizar)	Beckey	1969
Espectrómetro Triple-cuádruple	Yost and Enke	1978
Bombardeo rápido de átomos para su ionización *Fast Atom Borbardment (FAB)*	Barber	1981

Tabla 2.1: *Avances en la instrumentación referentes a la espectrometría de masas*

El análisis de metabolitos depende en gran medida de los avances tecnológicos y computacionales en la instrumentación analítica y la sensibilidad de dicha instrumentación. Por esta razón, los avances en la instrumentación de la espectrometría de masas tienen relevancia sobre los avances en la metabolómica.

Los estudios de genómica, cuyo mayor avance se dio en la década de 1980, y proteómica, principalmente desarrollada en la de 1990, comienzan a mostrar limitaciones a partir de la década de los 2000 a la hora de buscar marcadores para el estudio de las células. A partir de ese momento comienza a cobrar cada vez más importancia la metabolómica como ciencia clave para comprender las alteraciones celulares que se sufren en la enfermedad.

Este pequeño repaso a la historia de la metabolómica y la instrumentación utilizada para su estudio nos lleva al momento actual, caracterizado por la disponibilidad de instrumentación experimental avanzada y de alta resolución. En la actualidad los estudios metabolómicos ya no se ven limitados principalmente por la funcionalidad y precisión de los espectrómetros de masas. El principal cuello de botella en estos estudios está en la subsecuente interpretación de los datos obtenidos por la instrumentación. Sigue siendo habitual el descubrir señales provenientes de metabolitos cuyas masas moleculares no pueden ser identificadas con ningún

metabolito descrito en las rutas metabólicas conocidas hasta el momento. En este proceso de identificación de metabolitos juegan un papel fundamental las distintas bases de datos de metabolitos disponibles en la actualidad.

2.2. Bases de datos metabolómicas

Uno de los actuales proyectos destinado a la investigación metabolómica es el *Human Metabolome Project* (HMP), desarrollado en la Universidad de Alberta, y dirigido por David Wishart[11]. Este proyecto, financiado con 7,5 millones de dólares por la organización *Genome Canada*, fue lanzado en 2005 con el propósito de mejorar la identificación, descubrimiento y la supervisión de enfermedades a partir de estudios metabolómicos. En él se desarrollan diferentes herramientas de software metabolómicas. El principal objetivo del proyecto es el descubrimiento de nuevos metabolitos que pueden ser encontrados en los tejidos y biofluidos del cuerpo humano con concentraciones mayores de un micromol. Los datos que se averiguan están disponibles en su base de datos, la *Human Metabolome Database*[41,65,66,68]. Este proyecto tiene como principal objetivo el proporcionar información de estos metabolitos, pero no está dedicado, por falta de presupuesto y recursos, a la explotación de los resultados para el procesado de los datos para identificación de nuevas enfermedades. Sí facilitan información sobre concentraciones de metabolitos asociados a enfermedades ya conocidas. Otro de los objetivos de este proyecto es combinar la información de los datos obtenidos a través de diferentes técnicas aplicadas a las muestras en el espectrómetro de masas (LC/MS, GC/MS o CE/MS) con la información de NMR para obtener mayor conocimiento acerca de la estructura de los metabolitos y sus rutas metabólicas.

The Scripps Research Institute (TSRI), organización privada sin ánimo de lucro donde existe un departamento dedicado a la metabolómica, es otro proyecto que está investigando actualmente este campo. Al contar con mayor cantidad de recursos que el HMP, en este proyecto no sólo tratan de descubrir nuevos metabolitos para ayudar a los estudios metabolómicos, sino que también desarrollan herramientas para el procesamiento de los datos. Tienen

una base de datos propia llamada Metlin[32,55], una herramienta alojada en la nube para el procesamiento informático de información procedente de estudios metabolómicos y herramientas para utilizar espectrometría de masas a nanoestructuras (*Nanostructure-Initiator Mass Spectrometry*). En el caso de este proyecto facilitan también un servicio de procesado completo de muestras externas bajo demanda según el modo de ionización utilizado con un precio de 30 a 100 dólares por muestra si se desea una técnica ESI-TOF (Ionización por electro spray con tiempos de retención), de 70 a 600 dólares por muestra si la técnica que se quiere emplear para el análisis es MALDI-TOF (Técnica Maldi para ionización con tiempos de retención) y ESI-TOF, de 30 a 150 dólares si la técnica es LC/QQQ (triple cuádruple, espectrometría en tándem para identificación a partir de fragmentos) cuantitiva, Q-TOF (cuádruple, es decir, incluyendo espectrometría en tándem, con tiempos de retención) o GC en sus dos variantes: GC/APCI/Q-TOF (creación de iones a temperatura atmosférica aplicando espectrometría en tándem con tiempos de retención) o GC/MS (cromatografía de gases sin aplicar espectrometría en tándem). Al ser un proyecto que pertenece a una entidad privada el acceso a sus datos tiene mayores restricciones que los proyectos desarrollados con mayor afán investigador, pero muchas de sus herramientas sí son gratuitas y contribuyen a la comunidad científica con una extensa base de datos cuya información es muy valorada por los químicos analíticos.

El estado en que se encuentra ahora mismo la identificación de resultados generados mediante espectrometría de masas en cualquiera de las técnicas empleadas es embrionario. Existen múltiples bases de datos y aplicaciones donde pueden buscarse coincidencias en función de los datos del espectro, pero todas las herramientas utilizan su base de datos propia y en la mayoría de los casos no ofrecen procesamiento posterior. Sí hay herramientas software para el completo procesamiento del espectro como MzMatch o MetAssign, pero parten de un filtrado que hay que realizar previamente, o utilizan una única base de datos para dicha identificación. Esta es una lista de las principales bases de datos disponibles para la identificación de metabolitos:

- Metlin.
- LipidMaps.
- KEGG.
- mzCloud.
- Mine.
- HMDB.

Las características de las mismas difieren en función de la aplicación. Metlin[32] es probablemente la más utilizada. Se ha especializado en la espectrometría de alta resolución en tándem y tiene actualmente más de 242.000 compuestos, de los cuales ofrecen datos acerca de espectrometría en tándem (**MS/MS**) de 72.268 y de alta resolución de 14.034[32]. Dicha información de MS/MS es obtenida mediante tecnología de Agilent (6510 Q-TOF) operando con técnicas de electro-spray en modo negativo y positivo y con diferente energía de colisión (0, 10, 20 y 40 V). LipidMaps[24] es una base de datos sólo de lípidos. Aporta información de los lípidos y los clasifica en función de su categoría, su clase principal y su clase secundaria. KEGG[40] (**K**yoto **E**ncyclopedia of **G**enes and **G**enomes) es una base de datos japonesa que aporta información de los metabolitos y de las rutas biológicas donde es conocido que se encuentran dichos metabolitos. También aporta información sobre reacciones en las que forma parte el metabolito y la enzima a la que pertenecen. La base de datos mzCloud ofrece información de alta resolución de metabolitos y sobre la espectrometría aplicada a dichos metabolitos. La característica más innovadora que ofrece es una técnica para la búsqueda de iones precursores de una molécula en caso de que dicha molécula no forme parte de la base de datos (Técnica *PIF*). Está diseñada para realizar búsquedas compuesto a compuesto pero no para búsqueda de compuestos en bloque. Mine ha desarrollado algoritmos para la aplicación de reglas químicas generales a los metabolitos conocidos para la generación de compuestos que no han sido previamente identificados. Esta característica hace que sea una

base de datos con interés en su integración con vistas a incluir conocimiento futuro sobre identificación de compuestos que no han sido previamente detectados. HMDB[41] recoge metabolitos presentes en los seres humanos y aporta información muy detallada del compuesto con multitud de referencias hacia otras bases de datos, aunque está comprobado por la experiencia de los químicos analíticos y del autor del proyecto que no toda la información es correcta. Un ejemplo de información incorrecta puede verse en la figura 2.1. En ella aparecen 378 compuestos que tienen una referencia al compuesto **C00626** de KEGG, compuesto que actualmente no existe en la fuente de datos, tal y como se puede apreciar en la figura 2.2.

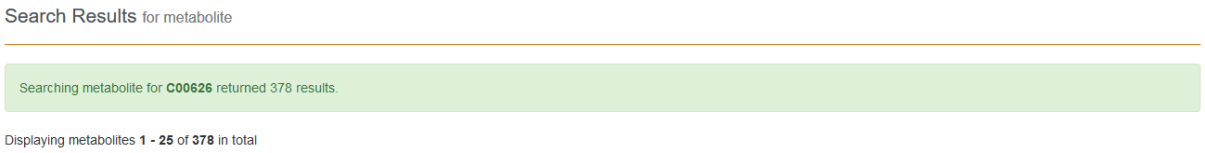


Figura 2.1: *Búsqueda de compuestos en HMDB con referencia al compuesto* ***C00626*** *de KEGG*[41]

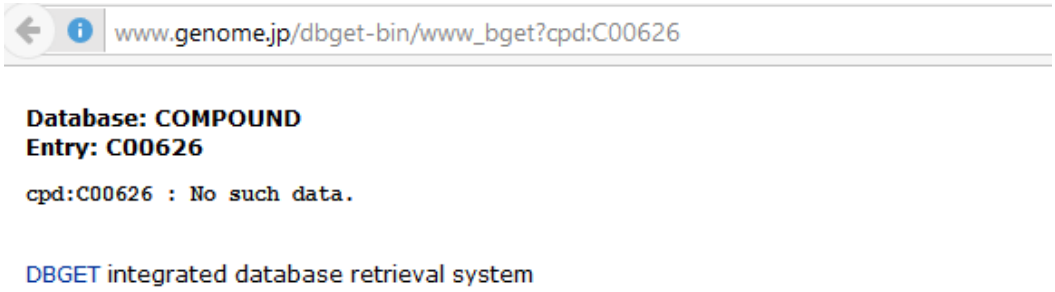


Figura 2.2: *Compuesto* ***C00626*** *en KEGG (Inexistente)*[40]

Lamentablemente la información disponible en estas bases de datos no siempre es ni correcta ni consistente. En el capítulo 4 se hace una descripción detallada de algunos de los problemas encontrados en las bases de datos que se han integrado.

En la tabla 2.2 se explica detalladamente la información facilitada por cada una de las bases de datos, haciendo un énfasis especial en aquella información que puede ser empleada para unificar compuestos. Cada fila muestra si un determinado dato es proporcionado o no

DB/Disponibilidad	KEGG	Metlin	LipidMaps	HMDB
Masa	Sí	Sí	Sí	Sí
Formula	Sí	Sí	Sí	Sí
Cas	Sí	Sí	No	Sí
InChIKey	No	No	Sí	Sí
InChI	No	No	Sí	Sí
Pathways	Sí	No	No	Sí
Fichero .mol	Sí	No	Sí	Sí

Tabla 2.2: *Información proporcionada por cada una de las fuentes*

por la base de datos correspondiente. La masa se refiere a la masa molécular del metabolito. La fórmula química del metabolito muestra los elementos que componen cada molécula. Esta información (masa molecular y fórmula química) no es suficiente por sí sola para la unificación de compuestos, pues existen multitud de metabolitos diferentes con misma fórmula y peso molecular. El **CAS Registry Number** (CAS) es un identificador mantenido por la sociedad americana de química que trata de identificar de forma unívoca al compuesto. El InChI es un identificador estándar para definir la estructura de los compuestos diseñado por la *International Union of Pure and Applied Chemistry* (**IUPAC**). Debido a su tamaño, a dicho identificador se le aplica un algoritmo hash para generar el InChIKey, identificador referente al InChI pero de menor tamaño. Las rutas metabólicas son rutas conocidas que se dan en las células. Las reacciones químicas que ocurren en la célula se agrupan y los metabolitos generados son a su vez anotados en diferentes rutas. El fichero .mol es uno de los formatos que existen para dibujar la estructura de los compuestos. A partir de estos ficheros puede calcularse la conectividad, y, en muchas ocasiones, generar el identificador InChI.

2.3. Herramientas para la identificación de compuestos

Existen varias herramientas que ayudan al análisis de datos procedentes del espectrómetro de masas. Estas herramientas por lo general trabajan con ficheros generados directamente

por parte del espectrómetro y procesan los datos realizando análisis estadísticos para buscar biomarcadores a partir del análisis de todas las muestras incluidas en el estudio. Una vez tienen caracterizados los metabolitos relevantes para el estudio, proceden a la identificación de los mismos, para lo que utilizan alguna de las bases de datos mencionadas anteriormente. Algunas de las principales herramientas son las siguientes:

- MzMatch.
- MetAssign.
- XCMS Online.
- MetAlign.
- Agilent MassHunter Profinder.

MzMatch[9] es una herramienta de código abierto desarrollada en Java y con un paquete disponible en R diseñado para el procesamiento de datos metabolómicos procedentes de muestras sobre las que se ha aplicado la técnica de cromatografía líquida (LC/MS). Está basado en ficheros con formato PeakML. Las características principales que facilita son la extracción de picos procedentes del espectrómetro, la combinación de diferentes ficheros .peakML[35,53], es decir, distintas muestras introducidas en el espectrómetro, para aplicar corrección de tiempos de retención y de masas atómicas. La herramienta realiza anotaciones sobre todas las filas, identificando los picos que pertenecen a los mismos metabolitos en origen, en función del tiempo de retención y de la masa. Para ello, aplica algoritmos para detectar correlaciones que indiquen que los picos corresponden a isótopos, aductos, multímeros o fragmentos de un metabolito precursor. Agrupa estos picos y realiza anotaciones sobre ellos para la posterior identificación en las bases de datos. Se pueden aplicar filtros a las señales, normalizarlas en función del fichero resultante de la combinación de experimentos, y tiene algoritmos de detección de picos relacionados correspondientes a un mismo metabolito. Por último, permite la utilización de bases de datos propias para la identificación

de metabolitos, pero estas bases de datos utilizan un formato propio. Existen herramientas para la conversión de ficheros proporcionados por la mayoría de bases de datos (Metlin no facilita sus bases de datos actualmente).

MetAssign[10] introduce un modelo de agrupación bayesiana (*clustering bayesiano*) para combinar información procedente del espectrómetro (Masa, RT e intensidad de los picos básicamente) para aumentar la eficacia en la anotación de picos. Puede ser usada esta anotación en el software MzMatch. En la figura 2.3 se puede apreciar la eficacia de MetAssign, MzMatch y Camera en función de una curva ROC cuyo eje horizontal indica el porcentaje de falsos positivos y en el eje vertical el porcentaje de verdaderos positivos. En dicha curva se puede apreciar que en función del ratio de identificaciones positivas o negativas que se quiera conseguir, puede ser más interesante el uso de MetAssign o MzMatch, pero la eficacia de Camera está lejos de las logradas por estos dos softwares.

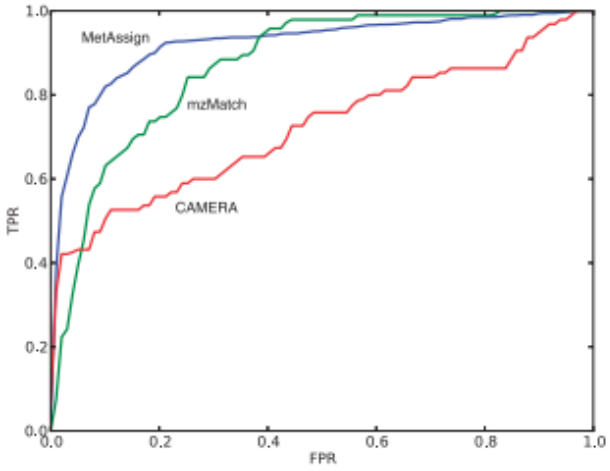


Figura 2.3: *Curva ROC perteneciente a diferentes software de anotación de metabolitos analizados mediante LC-MS*[10]

XCMS Online[61] ofrece la capacidad de crear trabajos para un procesamiento automático de los resultados con su herramienta XCMS Plus, que permite realizar análisis de rutas metabólicas, visualización de los datos utilizando diferentes modelos estadísticos en la búsqueda de marcadores y compartimento de datos. La búsqueda de metabolitos se realiza en

la base de datos de Metlin únicamente.

MetAlign[26] es un software para el pre-procesamiento y comparación de datos generados mediante cromatografía líquida o cromatografía de gases que se utiliza para realizar anotaciones sobre los resultados generados, así como la detección de contaminantes.

Agilent MassHunter Profinder no es exactamente una herramienta como el resto de las mencionadas aquí. Requiere de una licencia de pago que incorpora multitud de características para el análisis de datos y la búsqueda de los metabolitos relevantes, pero tiene más utilidades incorporadas como software para la puesta a punto del equipamiento que ofrecen para la espectrometría de masas. También facilita herramientas de filtrado para el procesamiento posterior y análisis recursivo sobre muestras, con algoritmos de puntuación sobre metabolitos y creación de centroides y perfiles mediante algoritmos de agrupamiento de diferentes características que pertenecen a un mismo metabolito (isótopos, aductos, multímeros, ...).

Existen otros softwares para la identificación de metabolitos como MolFind o Camera, pero los últimos estudios indican que la eficacia que tienen es menor que la de los descritos anteriormente en la figura 2.3[10]. Todas estas herramientas utilizan una única base de datos para la posible anotación de metabolitos, algo que puede limitar el volumen de metabolitos sobre los que se va a realizar la búsqueda. Además, no trabajan con posibles transformaciones en los metabolitos como la formación de aductos, ni permiten filtrar los resultados según los elementos que forman el compuesto (alfabeto químico).

La herramienta a desarrollar en este proyecto sería utilizada posteriormente a estas herramientas de análisis estadístico para conocer los posibles marcadores de un experimento. Las ventajas que aportaría esta herramienta son: la integración de diferentes bases de datos de búsqueda; la búsqueda automática de aductos; la utilización del espectro de composición para la detección de compuestos con doble carga; y el uso de un alfabeto químico. El análisis de rutas metabólicas sería un procesamiento posterior a la identificación que no tiene relación con estas herramientas.

2.4. Estado de Ceu Mass Mediator

La aplicación de la cual se partió en el desarrollo presentado en esta memoria consistía, por un lado, en un proyecto JavaSE para la actualización de base de datos y, por otro lado, de un proyecto JavaEE para la herramienta que hacía uso de dicha base de datos para la búsqueda de metabolitos. El proyecto estaba desarrollado sobre un servidor con un procesador Intel(R) Core(TM)2 Duo CPU E7200 con una frecuencia de 2.53GHz con 2Gb de memoria. El sistema operativo que utiliza es la distribución de Linux 3.1.0 OpenSUSE. En el servidor hay actualmente otras herramientas en producción, por lo que no es una tarea sencilla desde el punto de vista burocrático la modificación de la tecnología allí. La versión de Java que corría en el servidor es la OpenJDK 1.6.0_27. El servidor de aplicaciones donde estaba desplegado el proyecto es Glassfish 3.1.2, y los datos están en una base de datos MySQL 5.5.16. En el servidor existen otras aplicaciones usando esa base de datos, por lo que no es una opción actualizar la versión de MySQL.

La aplicación presentaba el modelo entidad-relación ya detallada en la figura 1.10. Los compuestos de diferentes fuentes eran tratados como entes independientes. La información procedente de las fuentes contiene numerosos fallos, por lo que los campos que se aprecian llamados **kegg_id**, **hmdb_id** o **metlin_id** en las diferentes tablas de compuestos no pueden emplearse para realizar una integración fiable de los datos procedentes de las distintas bases de datos. En este modelo entidad-relación se aprecian cuatro vistas (**metlin_compounds_view**, **lipid_maps_compounds_view**, **kegg_metlin** y **metlin_no_kegg**) y diez tablas. Todo el modelo está basado en una premisa que se ha demostrado equivocada. En teoría, el identificador de elementos químicos CAS es único y suficiente para la identificación y unificación de compuestos. Esta premisa es cierta desde un punto de vista teórico, pero la base de datos de CAS[58] es de acceso privado y no permite el acceso automático a sus datos para poder obtener la información real de cada número de registro patentado en cada metabolito. Así, por ejemplo, en la base de datos de HMDB hay 11 compuestos diferentes con un mismo identificador CAS para el CAS 71012-19-6, como se

aprecia en la figura 2.4, o los compuestos de KEGG **C19040** y **C08705**, los cuales tienen el mismo CAS y no tienen nada que ver entre ellos, como se aprecia en la figura 2.5

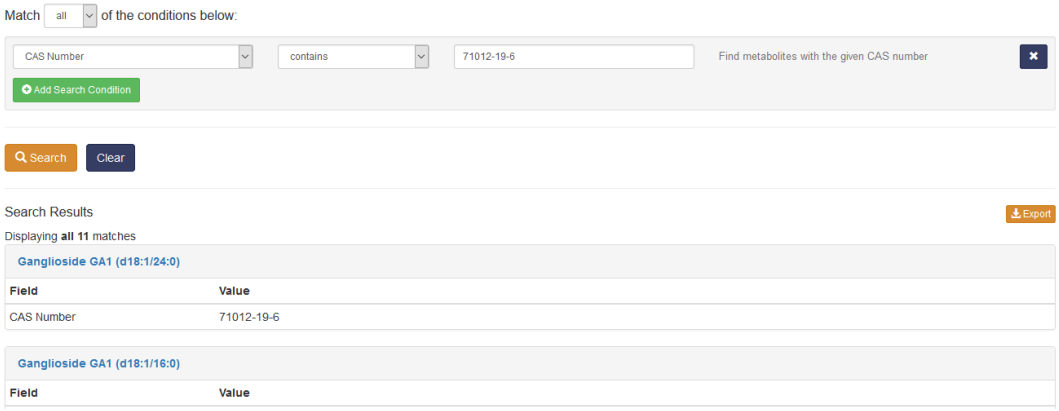


Figura 2.4: *Búsqueda de compuestos con identificador CAS 71012-19-6 en HMDB*[41]

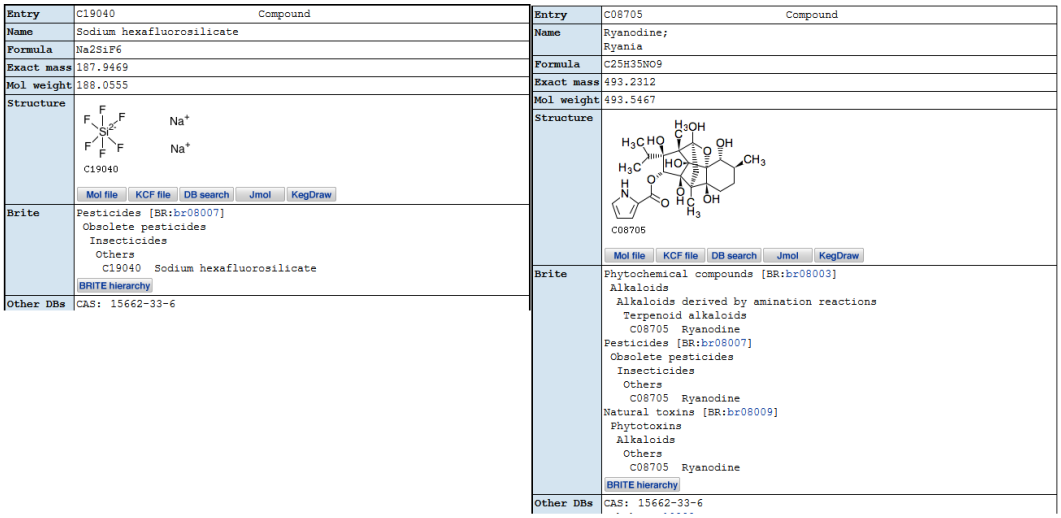


Figura 2.5: *Búsqueda de compuestos con identificador CAS 15662-33-6 en KEGG*[40]

En este modelo además se hacía uso de las tablas **fast_lipid_maps_compounds** y **fast_metlin_compounds** como una estrategia para disminuir los tiempos de acceso. En lugar de relacionar las tablas mediante el **kegg_id** para acceder a la columna **casIdentifier**, se creaba una tabla **metlin_compounds_view** y **lipid_maps_compounds_view** que contenía los campos de las tablas **metlin_compounds** y **lipid_maps_compounds** res-

pectivamente y la columna **casIdentifier** de la tabla **kegg_compounds** y se efectuaba un llenado de las tablas **fast_metlin_compounds** y **fast_lipid_maps_compounds** a partir de estas vistas. De este modo se evitaban las uniones (*joins*) que habría que realizar en cada consulta, pudiendo acceder a los datos directamente sobre una tabla y disminuyendo el tiempo de ejecución de la consulta. El problema de estas relaciones es que el CAS y las unificaciones proporcionadas por las fuentes no es fiable, tal y como se probó durante el proyecto. Por ejemplo, un mismo compuesto de KEGG estaba relacionado con 928 compuestos de LipidMaps, o 12 compuestos de HMDB tenían el mismo CAS.

Para la capa de presentación se utilizaba Java Server Pages y XHTML que se comunicaba con el back-end mediante JSF.

Capítulo 3

Diseño de la herramienta

Tras explicar las deficiencias que existen en las herramientas para la identificación de metabolitos y las aportaciones que se van a hacer este proyecto, en este capítulo se va a realizar el estudio de las necesidades del proyecto para proceder al diseño de la herramienta deseada. Tras desarrollar un análisis de requisitos junto a los responsables del CEMBIO (ver sección 3.1), se realizó un estudio sobre el estado previo de la herramienta (ver sección 2.4), y, posteriormente, se diseñó un nuevo modelo de datos para la aplicación (ver sección 3.3.3). También se desarrollaron una serie de prototipos (*sketchs*) (ver sección 3.2) para validar con todas las personas implicadas que el proyecto iba a cumplir con los objetivos iniciales y se realizó el estudio de las tecnologías que se utilizaron para desarrollar el proyecto.

3.1. Análisis de requisitos

En el caso concreto de este proyecto, los requisitos se establecieron mediante el método de entrevista con las personas implicadas. Las historias de usuario que se documentaron están recogidas en la sección 3.1.1. Aunque la especificación de Scrum recomienda que las historias de usuario sean cortas para poder modificarlas entre diferentes miembros del equipo, en este proyecto se han realizado historias de usuario más detalladas, debido a que el equipo de trabajo estaba formada por el autor. De esta manera, se obtenía una información mucho más detallada para su posterior implementación. Tras una serie de reuniones con los encargados del CEMBIO se decidieron los siguientes requisitos:

3.1.1. Requisitos funcionales

Tras las reuniones con los usuarios finales de la herramienta se identificaron ocho historias de usuario principales:

1. Búsqueda simple de metabolitos a partir una única masa experimental.
2. Búsqueda avanzada de metabolitos a partir de una única masa experimental.
3. Búsqueda simple de metabolitos a partir de un conjunto de masas experimentales.
4. Búsqueda avanzada de metabolitos a partir de un conjunto de masas experimentales.
5. Listado de resultados.
6. Generación de ficheros .xls para búsqueda de metabolitos.
7. Análisis rutas metabólicas.
8. Generación de ficheros .xls a partir del análisis de rutas metabólicas.

En las tablas 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7 y 3.8 se muestra en detalle cada una de estas historias de usuario.

Nombre	Búsqueda simple de una masa experimental.
Versión	3.0 (02/06/2016)
Dependencias	No Aplica
Precondiciones	1. Masa experimental introducida por el usuario. 2. Tolerancia introducida por el usuario. 3. Modo de masa experimental introducida por el usuario. 4. Modo de ionización introducido por el usuario. 5. Posibles aductos introducidos por el usuario.
Descripción	El sistema debe devolver los resultados al buscar entre las bases de datos integradas la masa experimental introducidas por el usuario en función de la tolerancia admitida.

Secuencia	1. El usuario introduce la masa experimental. 2. El usuario introduce la tolerancia permitida para la búsqueda de metabolitos en las bases de datos. 3. El usuario introduce el tipo de masa experimental, (neutras, m/z o masa m/z recalculadas). 4. El usuario introduce el modo de ionización (neutro, positivo o negativo). 5. Se realiza la búsqueda en la base de datos de Ceu Mass Mediator. 6. Se muestran los resultados de la búsqueda, posibles metabolitos correspondientes a las masa experimental introducida. Los resultados están ordenados en función de la masa experimental y la búsqueda por aductos.
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito con un peso igual a las masa experimental y la tolerancia introducidas por el usuario, el sistema informará al usuario de que la consulta no ha devuelto resultados.
Variantes	▪ El usuario puede escoger las fuentes de datos donde van a buscarse los posibles metabolitos correspondientes a la masa experimental. ▪ El usuario puede escoger los diferentes aductos que han podido formarse durante el proceso de análisis en el espectrómetro de masa.
Comentarios	▪ El valor de la tolerancia debe estar en el intervalo (0-1.000]. ▪ El número máximo de metabolitos a buscar es de 1.000.

Tabla 3.1: *Historia de usuario correspondiente a la búsqueda simple de metabolitos a partir una única masa experimental*

Nombre	Búsqueda avanzada de una masa experimental.
Versión	3.0 (02/06/2016)
Dependencias	Extiende búsqueda simple
Precondiciones	1. Masa experimental introducida por el usuario. 2. Tolerancia introducida por el usuario. 3. Tiempo de retención introducido por el usuario. 4. Composición del espectro introducido por el usuario. 5. Alfabeto químico introducido por el usuario. 6. Modo de masa experimental introducida por el usuario. 7. Modo de ionización introducido por el usuario. 8. Posibles aductos introducidos por el usuario.
Descripción	El sistema debe devolver los resultados al buscar entre las bases de datos integradas las masa experimental introducida por el usuario en función de la tolerancia admitida.

Secuencia	1. El usuario introduce la masa experimental. 2. El usuario introduce la tolerancia permitida para la búsqueda de metabolitos en las bases de datos. 3. El usuario introduce el tiempo de rentención. 4. El usuario introduce la composición del espectro. 5. El usuario introduce el alfabeto químico de la muestra. 6. El usuario introduce el tipo de de entrada de la masa experimental (neutras, m/z o masa m/z recalculadas). 7. El usuario introduce el modo de ionización (neutro, positivo o negativo). 8. Se realiza la búsqueda en la base de datos de Ceu Mass Mediator. 9. Se muestran los resultados de la búsqueda, posibles metabolitos correspondientes a la masa experimental introducida. Los resultados están ordenados en función de la masa experimental y la búsqueda por aductos.
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito con un peso igual a la masa experimental y la tolerancia introducida por el usuario, el sistema informará al usuario de que la consulta no ha devuelto resultados.
Variantes	▪ El usuario puede escoger las fuentes de datos donde van a buscarse los posibles metabolitos correspondientes a la masa experimental. ▪ El usuario puede escoger los diferentes aductos que han podido formarse durante el proceso de análisis en el espectrómetro de masa. ▪ El usuario puede escoger el alfabeto químico del compuesto entre CHNOPS, CHNOPS + Cl y todos los elementos químicos.

Comentarios	▪ El valor de la tolerancia debe estar en el intervalo (0-1.000]. ▪ El número máximo de metabolitos a buscar es de 1.000.

Tabla 3.2: *Historia de usuario correspondiente a la búsqueda avanzada de metabolitos a partir de una única masa experimental*

Nombre	Búsqueda simple de un conjunto de masas experimentales.
Versión	3.0 (02/06/2016)
Dependencias	Incluye búsqueda simple
Precondiciones	1. Masas experimentales introducidas por el usuario. 2. Tolerancia introducida por el usuario. 3. Modo de masas experimentales introducida por el usuario. 4. Modo de ionización introducido por el usuario. 5. Posibles aductos introducidos por el usuario.
Descripción	El sistema debe devolver los resultados al buscar entre las bases de datos integradas las masas experimentales introducidas por el usuario en función de la tolerancia admitida.

Secuencia	1. El usuario introduce las masas experimentales. 2. El usuario introduce la tolerancia permitida para la búsqueda de metabolitos en las bases de datos. 3. El usuario introduce el tipo de masas experimentales, (neutras, m/z o masas m/z recalculadas). 4. El usuario introduce el modo de ionización (neutro, positivo o negativo). 5. Se realiza la búsqueda en la base de datos de Ceu Mass Mediator. 6. Se muestran los resultados de la búsqueda, posibles metabolitos correspondientes a las masas experimentales introducidas. Los resultados están ordenados en función de la masa experimental y la búsqueda por aductos.
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito con un peso igual a las masas experimentales y la tolerancia introducidas por el usuario, el sistema informará al usuario de que la consulta contiene resultados.
Variantes	▪ El usuario puede escoger las fuentes de datos donde van a buscarse los posibles metabolitos correspondientes a las masas experimentales. ▪ El usuario puede escoger los diferentes aductos que han podido formarse.
Comentarios	▪ El valor de la tolerancia debe estar en el intervalo (0-1.000]. ▪ El usuario puede escoger los diferentes aductos que han podido formarse durante el proceso de análisis en el espectrómetro de masa.

Tabla 3.3: *Historia de usuario correspondiente a la búsqueda simple de metabolitos a partir de un conjunto de masas experimentales*

Nombre	Búsqueda avanzada de un conjunto de masas experimentales.
Versión	3.0 (02/06/2016)
Dependencias	Incluye búsqueda avanzada
Precondiciones	1. Masas experimentales introducidas por el usuario. 2. Tolerancia introducida por el usuario. 3. Tiempos de retención introducidos por el usuario. 4. Composición de espectros introducido por el usuario. 5. Alfabeto químico de la estructura introducido por el usuario. 6. Modo de masas experimentales introducida por el usuario. 7. Modo de ionización introducido por el usuario. 8. Posibles aductos introducidos por el usuario.
Descripción	El sistema debe devolver los resultados al buscar entre las bases de datos integradas las masas experimentales introducidas por el usuario en función de la tolerancia admitida.

Secuencia	1. El usuario introduce las masas experimentales. 2. El usuario introduce la tolerancia permitida para la búsqueda de metabolitos en las bases de datos. 3. El usuario introduce los tiempo de rentención. 4. El usuario introduce las composiciones del espectro para cada masa experimental. 5. El usuario introduce el alfabeto químico de la muestra. 6. El usuario introduce el tipo de masas experimentales, (neutras, m/z o masas m/z recalculadas). 7. El usuario introduce el modo de ionización (neutro, positivo o negativo). 8. Se realiza la búsqueda en la base de datos de Ceu Mass Mediator. 9. Se muestran los resultados de la búsqueda, posibles metabolitos correspondientes a las masas experimentales introducidas. Los resultados están ordenados en función de la masa experimental y la búsqueda por aductos.
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito con un peso igual a las masas experimentales y la tolerancia introducidas por el usuario, el sistema informará al usuario de que la consulta no ha devuelto resultados.

Variantes	▪ El usuario puede escoger las fuentes de datos donde van a buscarse los posibles metabolitos correspondientes a las masas experimentales. ▪ El usuario puede escoger los diferentes aductos que han podido formarse durante el proceso de análisis en el espectrómetro de masa. ▪ El usuario puede escoger el alfabeto químico del compuesto entre CHNOPS, CHNOPS + Cl y todos los elementos químicos.
Comentarios	▪ El valor de la tolerancia debe estar en el intervalo (0-1.000]. ▪ El número máximo de metabolitos a buscar es de 1.000.

Tabla 3.4: *Historia de usuario correspondiente a la búsqueda avanzada de metabolitos a partir de un conjunto de masas experimentales*

Nombre	Listado de resultados.
Versión	3.0 (02/06/2016)
Dependencias	No aplica
Precondiciones	1. Historia de usuario de búsqueda (cualquiera de los cuatro que se describen en las tablas 3.1, 3.2, 3.3 y 3.4) realizada previamente con éxito.
Descripción	El sistema muestra los resultados encontrados en cada una de las historias de usuario correspondientes a la búsqueda.

Secuencia	1. Se muestran resultados agrupados en función de masa experimental y aducto en una lista paginada para cada masa experimental introducida en la búsqueda. Contienen: ▪ Identificador ▪ Masa molecular ▪ Flag (Opcional) ▪ Diferencia (partes por millón) entre la masa experimental y el peso molecular del metabolito ▪ Cas ▪ Nombre ▪ Fórmula ▪ Identificador KEGG ▪ Identificador HMDB ▪ Identificador LipidMaps ▪ Identificador Metlin ▪ Identificador Pub Chemical ▪ Rutas metabolómicas 2. Los resultados pueden ordenarse por el usuario.
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito con un peso igual a las masas experimentales y la tolerancia introducidas por el usuario, el sistema informará al usuario de que la consulta no ha devuelto resultados.
Variantes	No aplica
Comentarios	No aplica

Tabla 3.5: *Historia de usuario correspondiente a mostrar los resultados obtenidos a partir de las diferentes búsquedas*

Nombre	Generación de fichero .xls.
Versión	3.0 (02/06/2016)
Dependencias	Extiende listado de resultados
Precondiciones	1. Historia de usuario de listado de resultados 3.5 realizada previamente con éxito.

Descripción	El sistema muestra los resultados encontrados en cada uno de las historias de usuario correspondientes a la búsqueda en formato .xls.
Secuencia	1. El usuario pulsa el botón para generar el fichero en formato .xls 2. Se muestran resultados en un fichero .xls ordenados por masa experimental, aducto correspondiente y diferencia respecto a la masa experimental. Las columnas que describen cada compuesto son: ▪ Masa experimental ▪ Flag (Opcional) ▪ Identificador ▪ Masa molecular ▪ Diferencia (partes por millón) entre la masa experimental y el peso molecular del metabolito ▪ Aducto ▪ Nombre ▪ Fórmula ▪ Cas ▪ Identificador KEGG ▪ Identificador HMDB ▪ Identificador LipidMaps ▪ Identificador Metlin ▪ Identificador Pub Chemical ▪ Rutas metabolómicas
Poscondición	No Aplica
Excepciones	No aplica
Variantes	No aplica
Comentarios	No aplica

Tabla 3.6: *Historia de usuario correspondiente a la generación de ficheros .xls para búsqueda de metabolitos*

Nombre	Agrupación por rutas metabólicas.
Versión	3.0 (02/06/2016)
Dependencias	No Aplica
Precondiciones	1. Fichero .xls subido por el usuario con la siguiente estructura: ▪ La primera fila del fichero será ignorada ▪ La segunda fila del fichero contendrá las cabeceras ▪ Las cabeceras deben tener los nombres: • Experimental mass • Flag (Opcional) • Identifier • Molecular Weight • PPM Error • Adduct • Formula • Cas • KEGG • HMDB • LipidMaps • Metlin • PubChem • Pathways 2. El orden de las columnas no altera el resultado de la agrupación 3. Si alguna de las cabeceras no apareciese, dicho campo no sería procesado 4. La última columna debe ser la de rutas metabólicas, y a partir de esta cada ruta irá en las siguientes columnas
Descripción	El sistema debe devolver los metabolitos agrupados en función de las posibles rutas donde aparecen Los metabolitos que no están categorizados en ninguna ruta no aparecen en dicha agrupación.

Secuencia	1. El usuario carga el fichero .xls en la herramienta 2. La herramienta agrupa los compuestos por el código de la ruta metabolómica 3. Se muestran los resultados de los compuestos agrupados y se ordena por número de ocurrencias
Poscondición	No Aplica
Excepciones	Si no hay ningún metabolito categorizado por alguna ruta se informará al usuario de que no hay compuestos categorizados por ninguna de ellas.
Variantes	Estructura del fichero .xls ▪ El orden de las columnas no altera el resultado de la agrupación.
Comentarios	No aplica

Tabla 3.7: *Historia de usuario correspondiente al análisis de rutas metabólicas*

Nombre	Generación de fichero .xls para compuestos agrupados por rutas metabólicas.
Versión	3.0 (02/06/2016)
Dependencias	Extiende agrupación por rutas metabólicas
Precondiciones	1. Historia de usuario de agrupación por rutas metabólicas mostrado en la tabla 3.7 realizada previamente con éxito.
Descripción	El sistema muestra los compuesto agrupados por ruta metabolómica encontrados en cada una de las historias de usuario correspondientes a la búsqueda en formato .xls.

Secuencia	1. El usuario pulsa el botón para generar el fichero en formato .xls 2. Se muestran resultados en un fichero .xls ordenados por ruta metabolómica. Cada ruta metabolómica contiene los compuestos que están presentes en dicha ruta y en los datos de entrada y cada compuesto ocupa una nueva línea. Las columnas que describen cada compuesto son: ▪ Masa experimental ▪ Flag (Opcional) ▪ Identificador ▪ Masa molecular ▪ Diferencia (partes por millón) entre la masa experimental y el peso molecular del metabolito ▪ Aducto ▪ Nombre ▪ Fórmula ▪ Cas ▪ Identificador KEGG ▪ Identificador HMDB ▪ Identificador LipidMaps ▪ Identificador Metlin ▪ Identificador Pub Chemical
Poscondición	No Aplica
Excepciones	No aplica
Variantes	No aplica
Comentarios	No aplica

Tabla 3.8: *Historia de usuario correspondiente a la generación de ficheros .xls a partir del análisis de rutas metabólicas*

Además de estas historias de usuario se desea integrar una nueva base de datos entre las fuentes de datos que emplea la aplicación: HMDB.

3.1.2. Requisitos no funcionales

Los requisitos no funcionales del proyecto son los siguientes:

- Se debe emplear el servidor previamente utilizado para el alojamiento de los servidores de aplicación y de base de datos, pues el próximo cambio que se haga en cuanto a hardware deberá realizarse con un estudio previo con la alternativa de tener la aplicación en algún servicio de computación en la nube o un nuevo servidor físico.
- Preparación del modelo de datos para la inclusión de conocimiento que se usará en el futuro.

3.2. Diseño de la interfaz de usuario

El cliente (y, al mismo tiempo, usuario) de la herramienta tenía interés en modernizar la interfaz web disponible inicialmente. Una de las posibilidades para diseñar interfaces web con la participación y aprobación del usuario es la realización de prototipos que son escogidos, modificados y aprobados por todos los miembros involucrados y sin emplear tiempo en la implementación del código.

Para la realización del diseño de la interfaz web se utilizó el software Balsamiq[2]. Era deseable la aceptación previa de los responsables del laboratorio sobre la visualización final de la página. El sistema de prototipados es una buena alternativa para que el cliente, en este caso el CEMBIO, mostrase su opinión y sugerencias en la fase más temprana posible, para que la variación tenga el menor coste posible. Inicialmente el equipo de desarrollo creó una versión inicial de los prototipos de diseño de la aplicación. Esta versión se enseñó durante una reunión a los investigadores del CEMBIO, quienes aportaron sus ideas y pidieron algunos cambios sobre el diseño inicial. Los prototipos de diseño que se muestran aquí son el resultado de haber llevado a cabo estos cambios.

En la figura 3.1 se puede ver la página principal de este prototipo. La herramienta incluye un logo principal, un menú vertical en la cabecera, un sistema de acceso para usuarios

registrados y un pie de página con la información tanto de la herramienta como del proyecto en el que esta encuadrado este proyecto (laboratorio CEMBIO). Estas características estarán presentes en todas las páginas de la herramienta. Además, la página principal contiene una breve descripción de la herramienta.

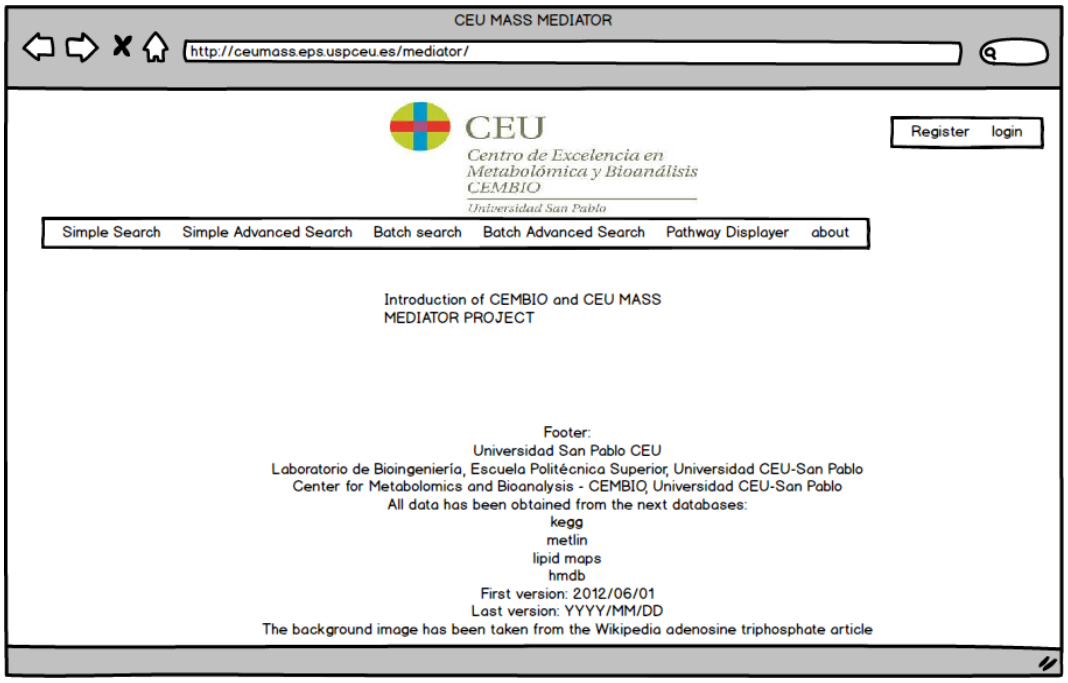


Figura 3.1: *Página principal del prototipo desarrollado*

En las figuras 3.2, 3.3, 3.4 y 3.5 se aprecian los prototipos de los diferentes tipos de búsqueda cuyas historias de usuario están descritos en las tablas 3.1, 3.2, 3.3 y 3.4. Estas búsquedas deben implementar todo lo descrito en su historia de usuario correspondiente. Hay una serie de información común para todas las búsquedas, como es la masa experimental, la tolerancia deseada para la búsqueda, el modo de la masa experimental, el modo de ionización y los posibles aductos que han podido ser formados durante el experimento. La búsqueda avanzada incluye además información sobre el tiempo de retención (posteriormente renombrado a *flag* por requerimiento del cliente), información acerca del espectro de composición formado para cada masa y el alfabeto químico. Esta información puede enviarse al servidor, que devolverá los resultados en el prototipo 3.6.

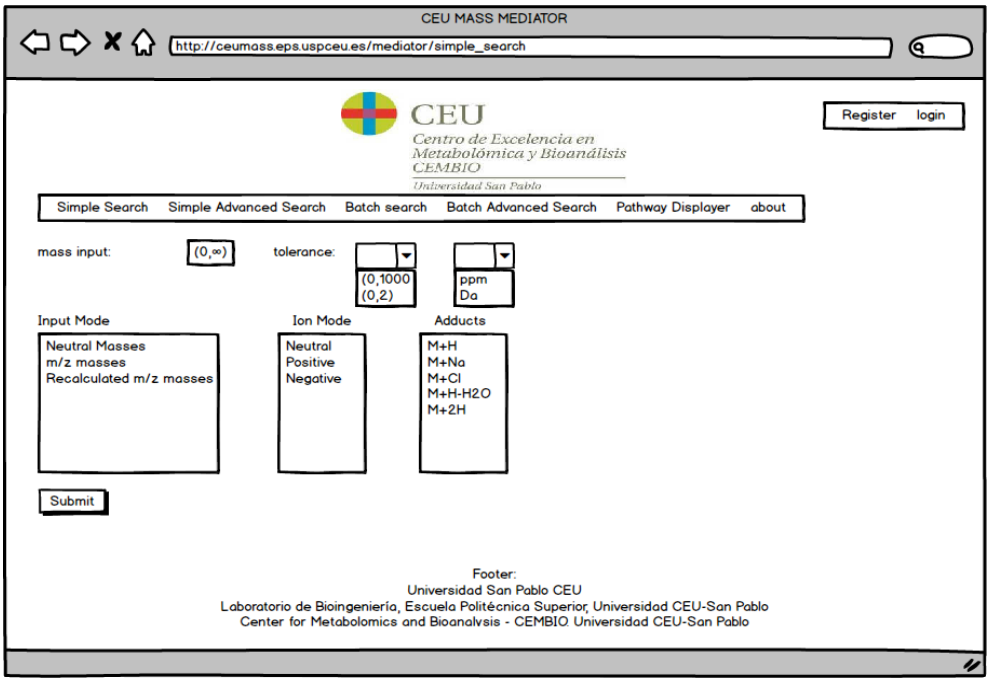


Figura 3.2: *Búsqueda simple*

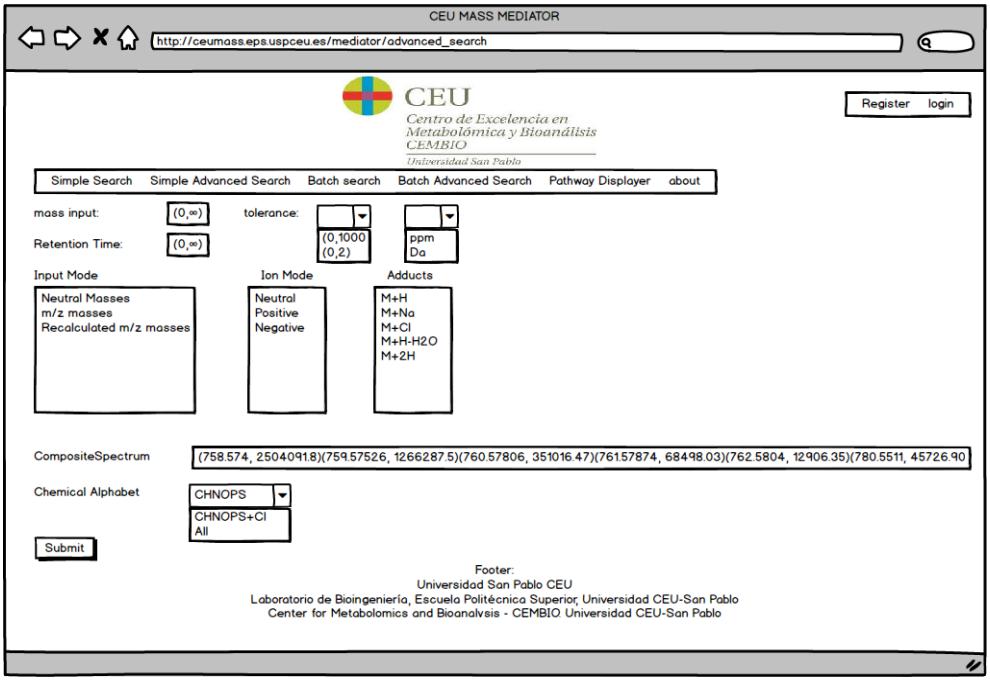


Figura 3.3: *Búsqueda avanzada*

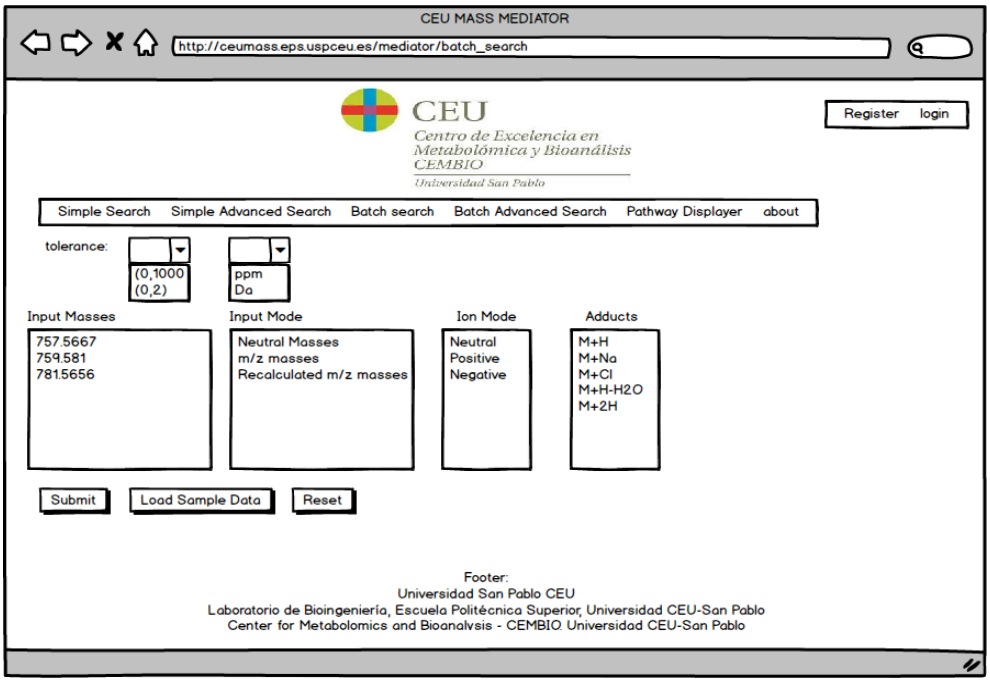


Figura 3.4: *Búsqueda múltiple simple*

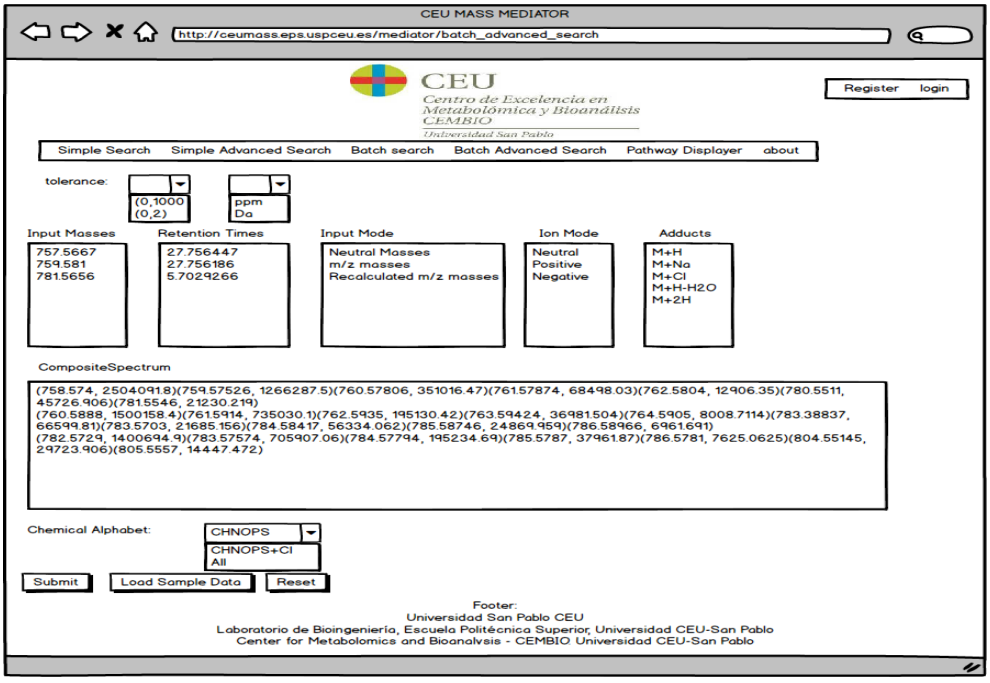


Figura 3.5: *Búsqueda múltiple avanzada*

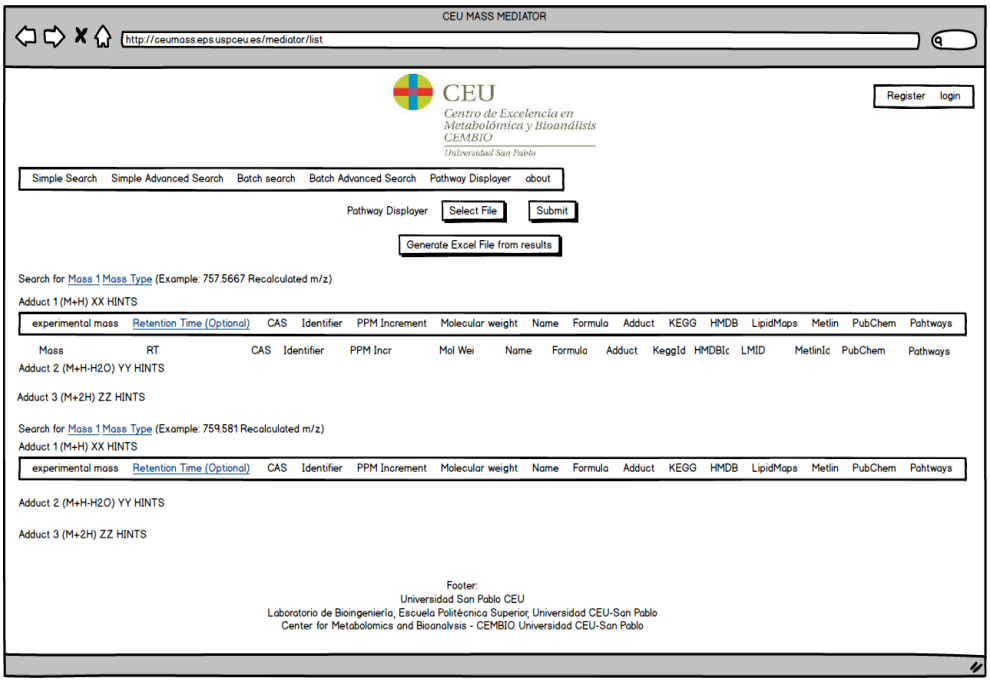


Figura 3.6: *Página de resultados*

Los prototipos correspondientes al análisis de rutas metabólicas, que incluyen la subida del fichero .xls con la información de los compuestos, se muestran en las figuras 3.7 y 3.8.

El diseño de la página de ayuda se puede ver en la figura 3.9 y el registro e inicio de sesión (*login*) en las figuras 3.10 y 3.11.

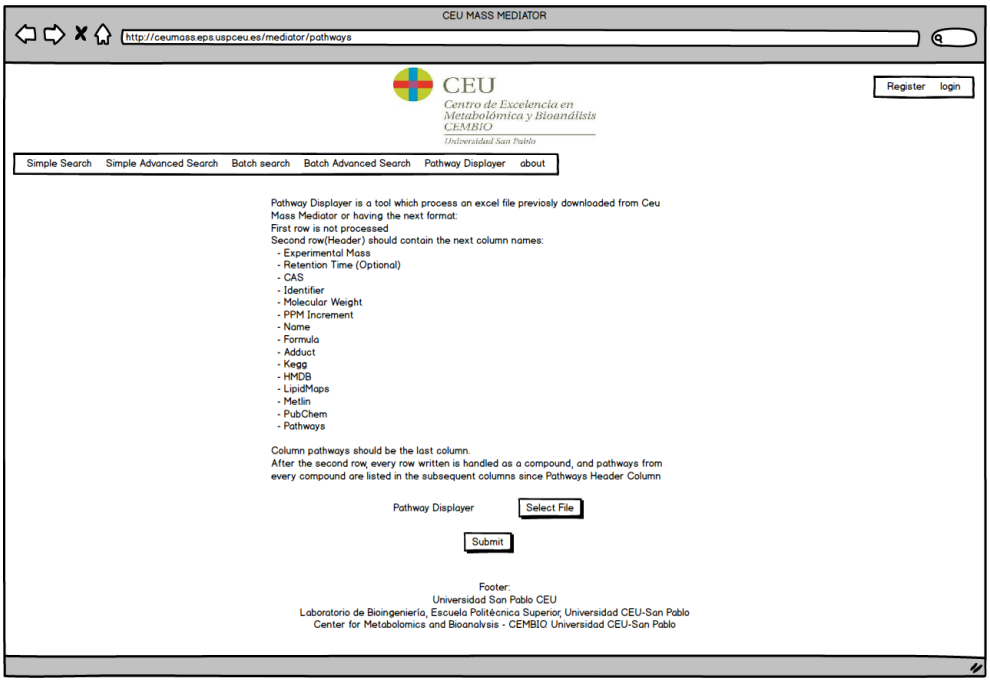


Figura 3.7: *Carga de compuestos para agrupación por rutas metabólicas*

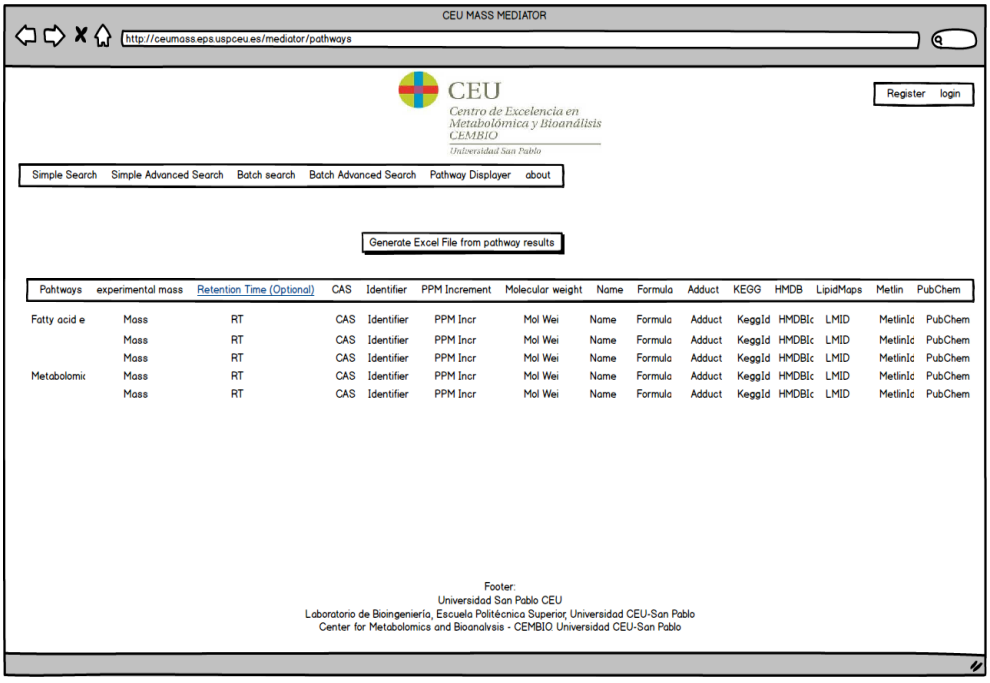


Figura 3.8: *Página de resultados de agrupación de compuestos por rutas metabólicas*

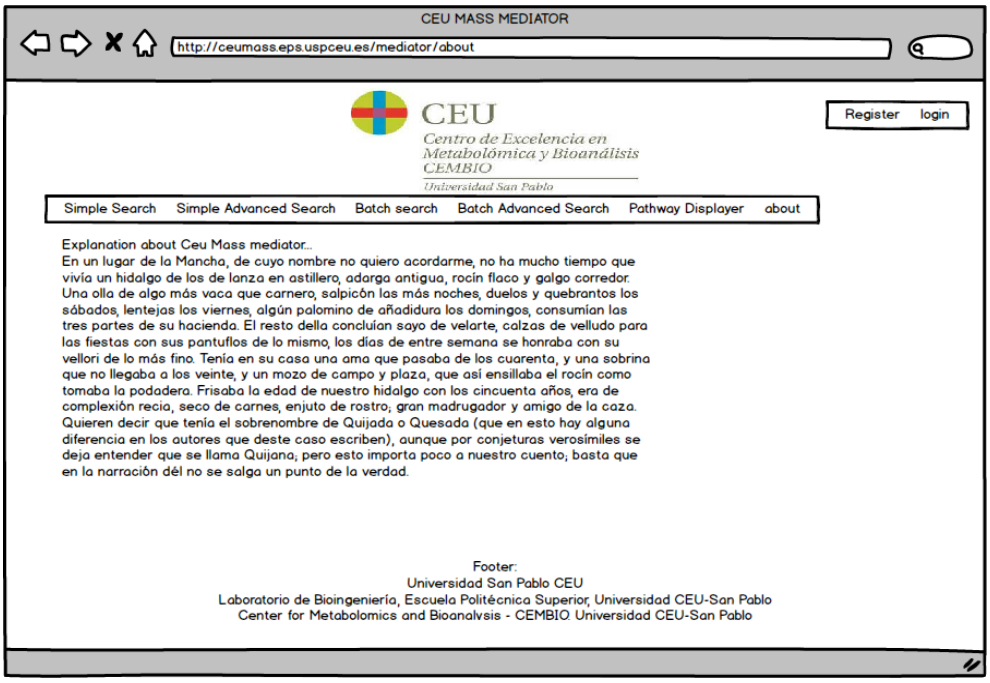


Figura 3.9: *Página de ayuda*

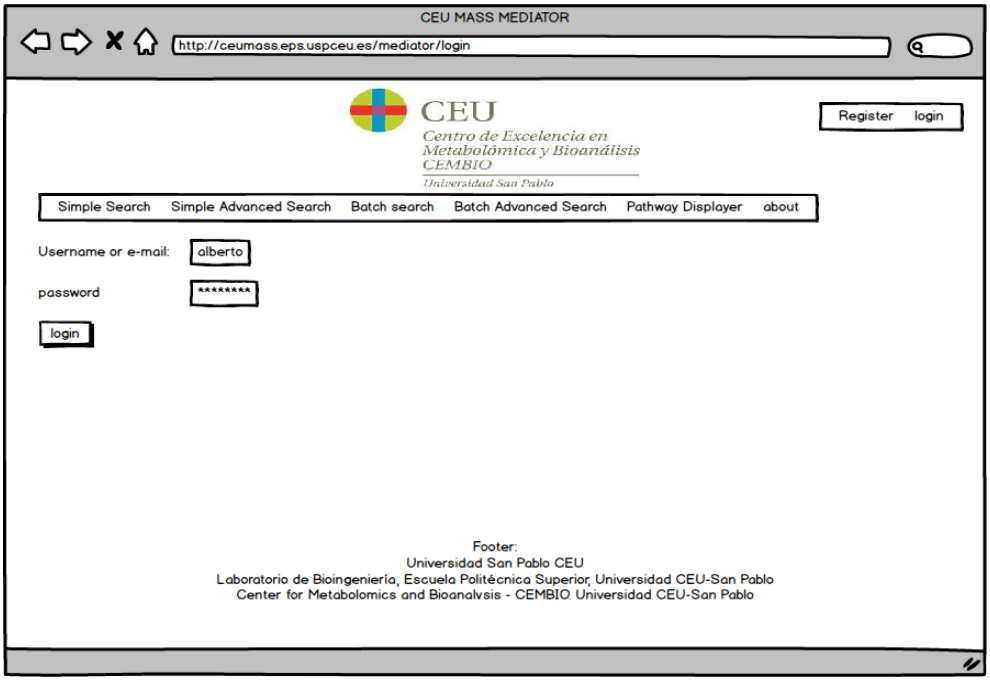


Figura 3.10: *Página de login*

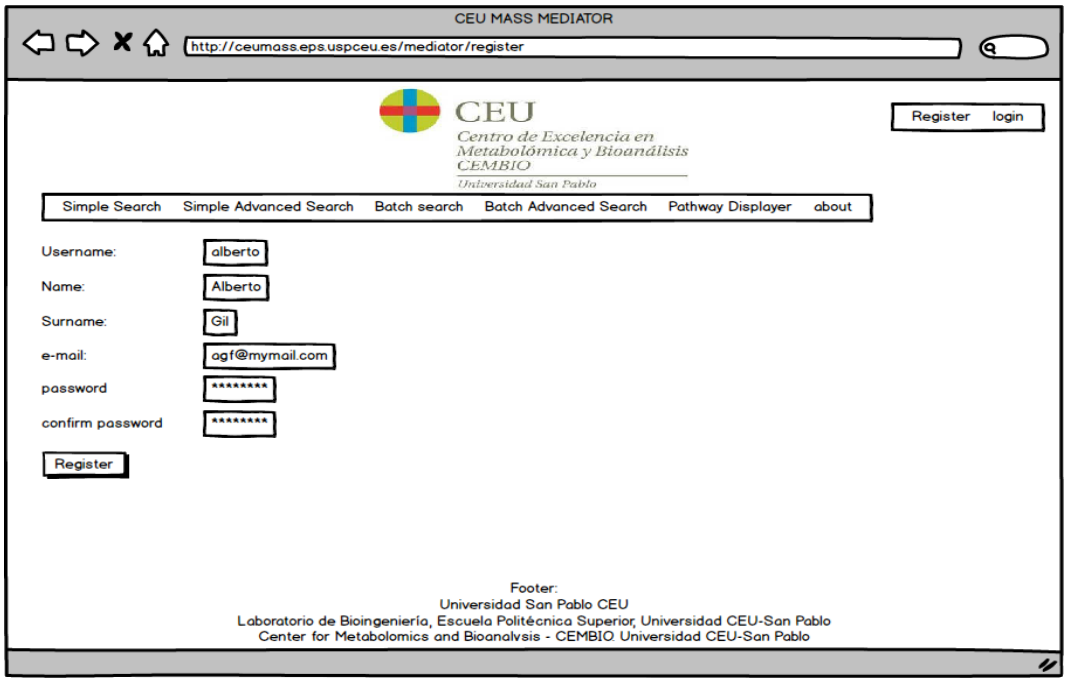


Figura 3.11: *Página de registro*

3.3. Tecnologías utilizadas

En este apartado se describirán la tecnología del cliente y del servidor que, tras un análisis de las distintas opciones disponibles, se decidieron emplear para desarrollar la aplicación. El único requisito a la hora de la elección de las tecnologías era que fuese útil para cumplir todos los requisitos funcionales y que preferiblemente la licencia fuese gratuita al tratarse de una aplicación de uso académico que quiere utilizarse para divulgación científica y tener un presupuesto limitado. En caso de que alguna tecnología de pago aportase una funcionalidad con un valor extra al proyecto se estudiaría su inclusión. Sin embargo, al no ser el caso en ninguno de los ámbitos evaluados, no se han utilizado licencias con coste.

3.3.1. Tecnología de desarrollo

La tecnología utilizada en el entorno de desarrollo ha sido Java Development Kit 8 (Java 8). Sin embargo, en el servidor de producción donde se alojará la aplicación, la versión insta-

lada era el Java Development Kit 6 (Java 6). Para la renovación de la interfaz web requerida por el cliente se actualizó la versión en el servidor a Java 8. Durante todo el desarrollo se ha utilizado el sistema de control de versiones Git[14] con el repositorio BitBucket[3] para facilitar el mantenimiento del código y tener un sistema de respaldo. Además, en caso de aumentar el equipo de desarrollo, Git proporciona funcionalidades para desarrollo no lineal, gestionando ramas que luego pueden mezclarse para su integración en la rama principal denominada "master". En el proyecto se han utilizado diferentes ramas a la hora de desarrollar nuevas funcionalidades, pero a la hora de integrar no ha habido cambios que unificar, puesto que había un único desarrollador.

Servidor

Se ha utilizado el JDK (*Java Development Kit*) versión 8, JavaEE 7 y el IDE (*Integrated Development Environment*) NetBeans. Se optó por usar Java porque esta tecnología es independiente de cualquier plataforma donde se utilice y facilita numerosas APIs para afrontar los problemas que se pretenden resolver con la herramienta. Se barajaron otras alternativas como .net, descartado al no aportar ninguna ventaja respecto a Java para la herramienta. Ademas, esta tecnología tiene el requisito de tener que funcionar bajo un sistema operativo Windows, y actualmente el servidor de producción no tiene este sistema operativo instalado. Ruby, Python y PHP mostraban características similares a Java, pero la experiencia del autor era menor en estas tecnologías, lo que hacía que el tiempo de desarrollo se viese incrementado si se utilizaban. Además, tanto el autor como el tutor tenían experiencia previa con el IDE mencionado.

En la aplicación se ha utilizado un Modelo-Vista-Controlador cuya lógica de presentación se desarrolló con el framework JSF 2.2 denominado FaceLets y la extensión de PrimeFaces[45], biblioteca que incluye funcionalidad avanzada basada en JavaScript que reduce el tiempo de desarrollo de la aplicación.

La lógica de negocio fue implementada mediante Enterprise Java Beans (EJB) de sesión que gestionan el flujo de información con la lógica de presentación y con el back-end. Se hizo

uso de Servlets para el procesado de peticiones http. Se ha hecho uso de la biblioteca Java Persistence API (JPA 2.0) para la persistencia de los datos, en concreto la implementación OpenJPA. Esta biblioteca permite tener una colección de objetos en memoria generados a partir de la información de la base de datos y persistir la información con la que se trabaje en la misma. La intercomunicación con la base de datos es administrada por las clases de JPA. La arquitectura de la aplicación en la figura 3.12.

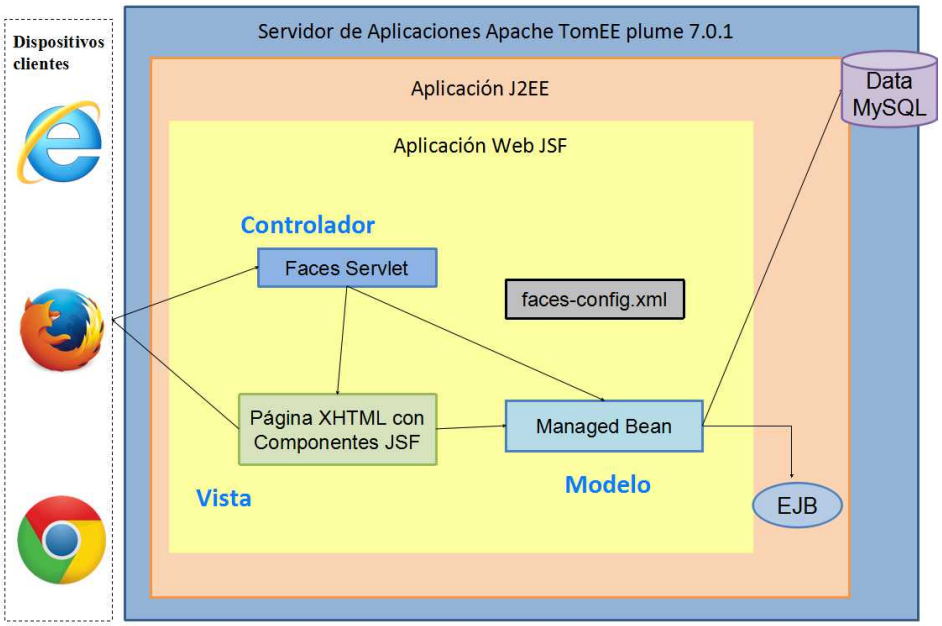


Figura 3.12: *Modelo modelo-vista-controlador de la aplicación*

Cliente

El cliente de la aplicación es el navegador web. Para la interfaz de usuario (Vista) se ha desarrollado código XHTML 1.1 con CSS3 y elementos de HTML5. Estas tecnologías permiten crear páginas adaptativas que se amolden a diferentes dispositivos como computadores tradicionales, tabletas o portátiles. Se ha hecho uso de composiciones y plantillas para aumentar la reusabilidad del código. Ha sido probado en Google Chrome, Internet Explorer, Mozilla Firefox y Opera, pero la aplicación es accesible desde cualquier navegador

web que soporte XHTML. Además de la portabilidad que aportan las aplicaciones web, al estar la herramienta enfocada para el uso por parte de químicos analíticos, habitualmente no muy acostumbrados a la configuración de aplicaciones, se ha preferido una aplicación web sobre una aplicación de escritorio, que hace necesaria la instalación en el equipo para su utilización.

3.3.2. Servidor de aplicaciones

El servidor de aplicaciones utilizado en la versión disponible inicialmente de la aplicación era Glassfish versión 3.1.2. Se estudió la posibilidad de actualizar a la última versión de Glassfish, pero éste tiene un fallo a la hora de crear nuevos *pools* de conexiones y está actualmente abandonado por Oracle. Por ello se descartó seguir usando este servidor de aplicaciones. La herramienta hace uso de la biblioteca PrimeFaces, que tiene funcionalidades que requieren una versión igual o superior a Java 7, versiones que no estaban instaladas en el servidor de producción. Por tanto, sería necesario actualizar la versión de Java del servidor. Tras analizar las posibilidades entre los diferentes servidores de aplicaciones se descartaron los sistemas propietarios (Oracle WebLogic, IBM WebSphere) y las alternativas barajadas fueron: JBoss, Apache Tomcat y Apache TomEE. Apache TomEE es una versión de Tomcat que, sin ser un servidor de aplicaciones Java EE completo, sí que cuenta con alguna funcionalidad de Java EE, como soporte para JSF o JPA. JBoss proporciona todas las funcionalidades de Java EE (OpenJPA, EJB, ...), al igual que TomEE, pero tras probar TomEE y comprobar su robustez y facilidad de uso, el autor se decidió por emplear este servidor de aplicaciones.

3.3.3. Base de Datos

El servidor de base de datos utilizado es MySQL. Era deseable tener un modelo de datos bien definido a la hora de unificar metabolitos procedentes de diferentes fuentes, por lo que se optó por emplear una base de datos relacional. MySQL es una base de datos relacional que actualmente pertenece a Oracle, pero que mantiene licencias gratuitas.

En el entorno de producción se utiliza la versión MySQL 5.5.16 y en el entorno de desarrollo se utiliza la versión 5.6.30 (recuérdese que por motivos ajenos a este proyecto no es posible actualizar la versión de la base de datos del servidor). No hay ninguna característica de la que se haga uso la aplicación que no esté disponible en la 5.5.16, y ésta cumplía con todos los requisitos del cliente, por lo que no se ha considerado necesaria la migración de la versión de base de datos en dicho entorno.

Modelo entidad-relación

Entre los objetivos de este proyecto se encontraba el unificar el modelo de datos empleado para representar los datos de cada una de las fuentes originales. En vez de contar con un conjunto de tablas diferentes para cada fuente, se pretendía unificar la información de todas estas fuentes en un único conjunto de tablas. Se deseaba tener compuestos con sus respectivas referencias a las diferentes bases de datos. El principal problema encontrado para la integración fue que, tras solicitar permiso al equipo de Metlin para acceder a sus datos e incluirlos en Ceu Mass Mediator, comunicaron que actualmente tienen la API fuera de servicio por motivos de seguridad. Era posible la utilización de los datos previamente accedidos pero no había posibilidad de actualizarlos. Por este motivo, los compuestos de Metlin han sido introducidos como entes diferentes y no han podido ser unificados. Se creó una entidad con la información de los metabolitos y diferentes entidades para cada fuente de datos con relación 1 a N sobre estos compuestos. Esta relación es 1:N y no 1:1 porque se han detectado metabolitos duplicados en las fuentes, en concreto 42 compuestos duplicados en KEGG y uno triplicado, en HMDB hay 99 compuestos duplicados y tres triplicados y en LipidMaps 28 compuestos duplicados. Cuando se habla de compuestos duplicados puede ser un mismo compuesto con dos referencias diferentes o realmente compuestos que son el mismo pero tienen dos entradas y no sólo dos referencias en las base de datos origen, dando lugar a incoherencias o falta de información en alguno de ellos. No se dispone de datos sobre Metlin acerca de compuestos duplicados ya que la información disponible sobre esta base de datos no permite la unificación de los metabolitos con total fiabilidad.

En la base de datos, existe una entidad para las rutas metabólicas, con una relación N:N con los compuestos. Al ir incluyendo conocimiento sobre los compuestos, éste será aplicable al compuesto como entidad, independientemente de la fuente a la que pertenezca.

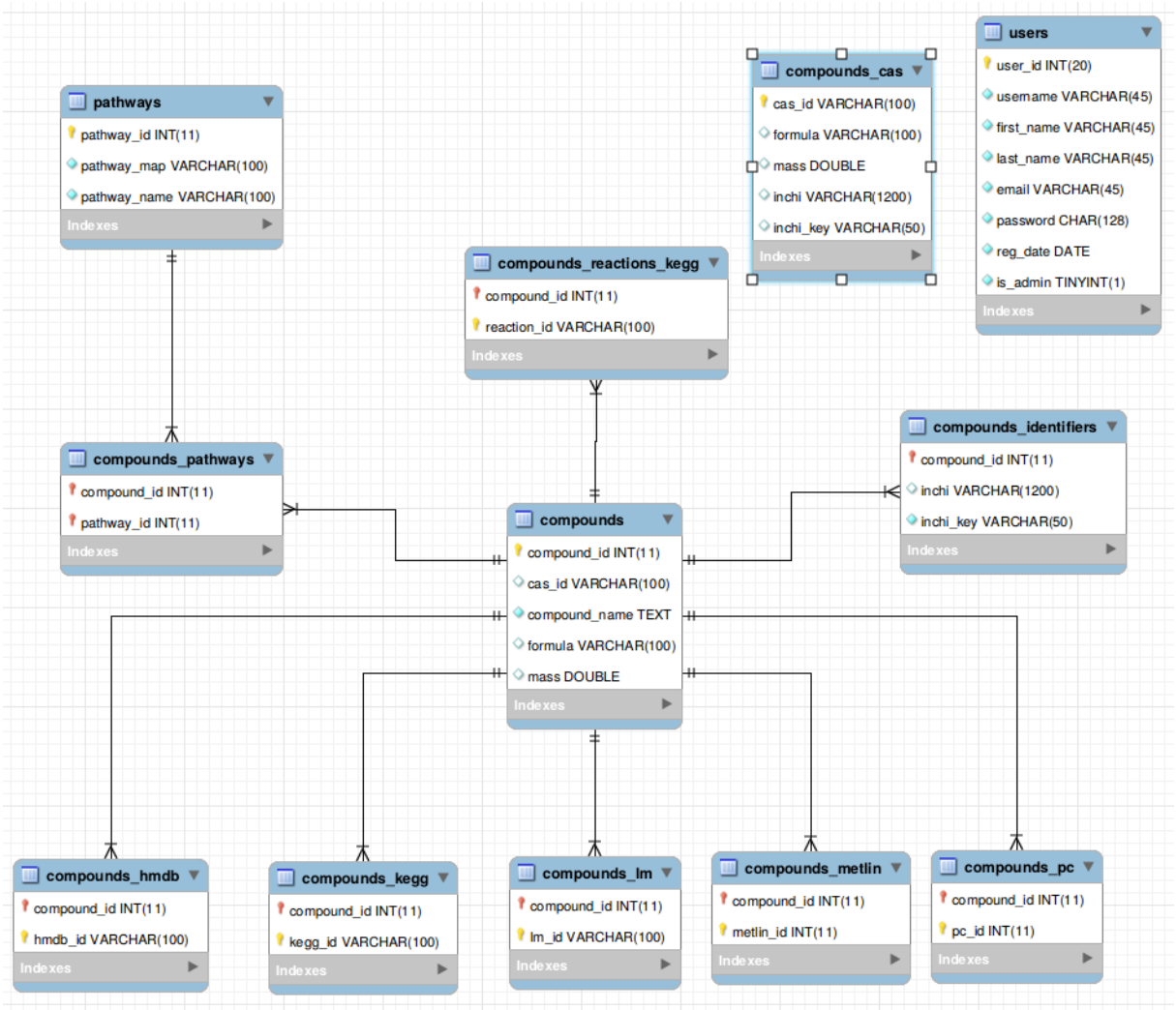


Figura 3.13: *Modelo entidad-relación de la aplicación desarrollada*

En la figura 3.13 se observa el nuevo modelo entidad-relación que está siendo usado actualmente por la herramienta. La entidad **compounds_identifiers** guarda el identificador único de cada compuesto, llamado **inchi_key**. A dicho identificador se le aplica un algoritmo hash para obtener su clave, guardada en la columna **inchi_key**. Hay una entidad denominada **compounds_cas** que guarda información acerca de los compuestos propor-

cionados por la Sociedad Americana de Química en su API oficial. La Sociedad Americana de Química es la encargada de registrar y mantener el CAS. Esta API no contiene todos los compuestos registrados, sólo algunos de ellos sin que aparentemente se siga ninguna regla sobre qué compuestos son proporcionados y cuáles no. Dicha entidad no tiene relación con los compuestos en el modelo entidad-relación, únicamente guarda la información oficial de los CAS. Existe una entidad para las posibles reacciones de donde proceden los compuestos que está incluida para un posible procesamiento futuro, pero actualmente no está siendo utilizada, que es la entidad **compounds_reactions_kegg**. Existe una entidad de usuarios para hacer una diferenciación en la herramienta entre usuarios con derechos de administrador, que en el futuro tendrán privilegios para añadir conocimiento, y usuarios comunes.

Capítulo 4

Implementación de back-end

En este capítulo se abordarán las cuestiones acerca de la implementación del back-end de la aplicación. En la sección 4.1 se hablará de los diferentes *sprints* de trabajo abordados durante el desarrollo. En la sección 4.2 se explicará cómo funciona la lógica de negocio de la herramienta. La sección 4.3 presenta la comunicación entre el back-end y la interfaz de presentación. La sección 4.4 desarrolla en profundidad todos los componentes que tienen que ver con las características del motor de la aplicación. Se contarán todos los detalles de la unificación de compuestos y los problemas encontrados durante dicha fase y por último se hablará de los *scripts* generados mediante tecnología Shell Script para la actualización automática de los datos de la aplicación, así como la generación de copias de seguridad de la base de datos mediante la herramienta de Crontab. El back-end se corresponde con la elipse roja dibujada en la figura 4.1.

4.1. *Sprints* de trabajo

La metodología Scrum (ver figura 1.13) explicada en la sección 1.5 estuvo formada por seis *sprints*, que se generaron en función de las necesidades del proyecto. El primer *sprint* de trabajo de la fase de implementación estuvo dedicado a la inclusión de búsquedas de aductos teniendo en cuenta el modo de ionización de los metabolitos (negativo, positivo o neutro). Al ser un proyecto perteneciente al CEMBIO, laboratorio que trabaja con espectrómetros de Agilent, se podía sacar provecho de algunas de las características que ofrecen

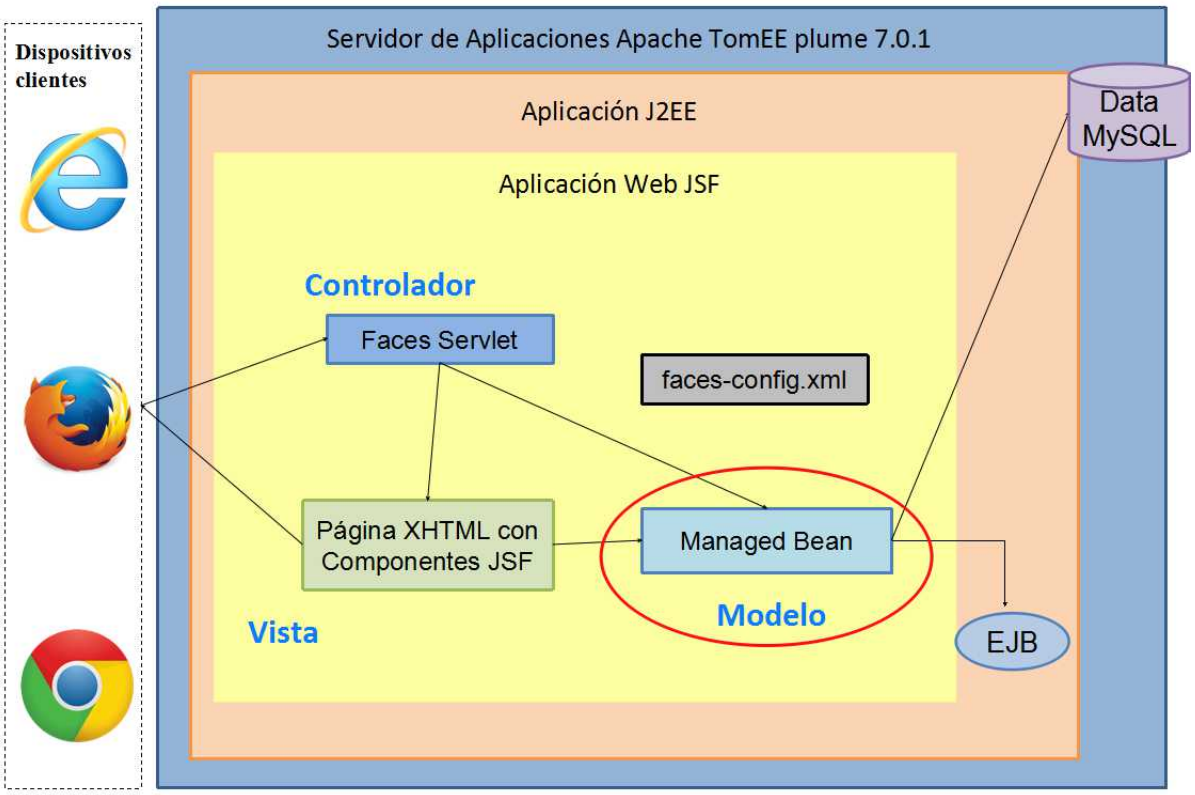


Figura 4.1: *Back-end en el modelo modelo-vista-controlador*

sus herramientas. Una de ellas es la detección de isótopos y aductos, que agrupa en la composición del espectro, pero que no funciona para el 100 % de los compuestos. Por lo tanto, es habitual tener que realizar la búsqueda de isótopos y aductos de forma manual. Para evitar esta búsqueda manual, en este primer *sprint* se realizó una búsqueda automática de aductos. Otra de las características es el tiempo de retención de cada compuesto. Mediante el tiempo de retención no puede obtenerse información relevante de las bases de datos, pues como se explicó en la introducción (capítulo 1), el tiempo de retención varía según condiciones del experimento y columna de separación. Por ello no es habitual que las bases de datos integradas incluyan información acerca del mismo. Sin embargo, la composición del espectro si es útil a la hora de detectar las cargas de cada una de las masas. A partir de esta composición pueden detectarse los aductos que tienen doble carga de forma automática, tal

y como se explica en la sección 4.4.

El segundo *sprint* de trabajo estuvo dirigido a la generación de ficheros .xls de acuerdo a los requisitos establecidos, tanto para la lista de resultados de posibles metabolitos identificados a partir de la masa experimental, como los resultados generados a partir de la agrupación por rutas metabólicas. Los requisitos están explicados en el caso de uso descrito en la tabla 3.2. El tercer *sprint* de trabajo estuvo dedicado a la integración de la base de datos HMDB. El cuarto *sprint* de trabajo se designó para la unificación de compuestos. El siguiente *sprint* de trabajo fue empleado en el análisis de rutas metabólicas. El sexto y último *sprint* de trabajo correspondiente al back-end se dedicó a la automatización de la actualización y a la creación de una política de copias de seguridad.

Durante todos los *sprints* de desarrollaron prototipos que iban siendo evaluados por el cliente y se iba recibiendo la opinión del mismo para realizar las modificaciones solicitadas o aprobar dichos prototipos. Una vez se alcanzó el visto bueno, se desarrolló e implantó el prototipo final. Al acabar cada uno de los *sprints* se realizó una revisión del mismo para probar si se cumplían los requisitos, recibir la confirmación del cliente y acordar el siguiente *sprint* en función de la prioridad de los requisitos aún por desarrollar.

4.2. Lógica de negocio

En esta sección se detallan las características de la lógica de negocio. Se define la lógica de negocio como la codificación de las reglas del mundo real que determinan el intercambio de información del usuario con la aplicación. La lógica de negocio recibe las peticiones del usuario, procesa dicha información y devuelve unos resultados que deben cumplir todas las reglas definidas en el diseño de la aplicación y aquí codificadas. En esta parte debe gestionarse la transferencia de información con el gestor de base de datos (descrito en el apartado 4.3), y con el usuario, a través de la capa de información, para recibir y procesar solicitudes. La arquitectura del modelo modelo-vista-controlador está desarrollada en la figura 4.2, y en ella puede verse la parte que corresponde a la lógica de negocio, que es la capa del controlador.

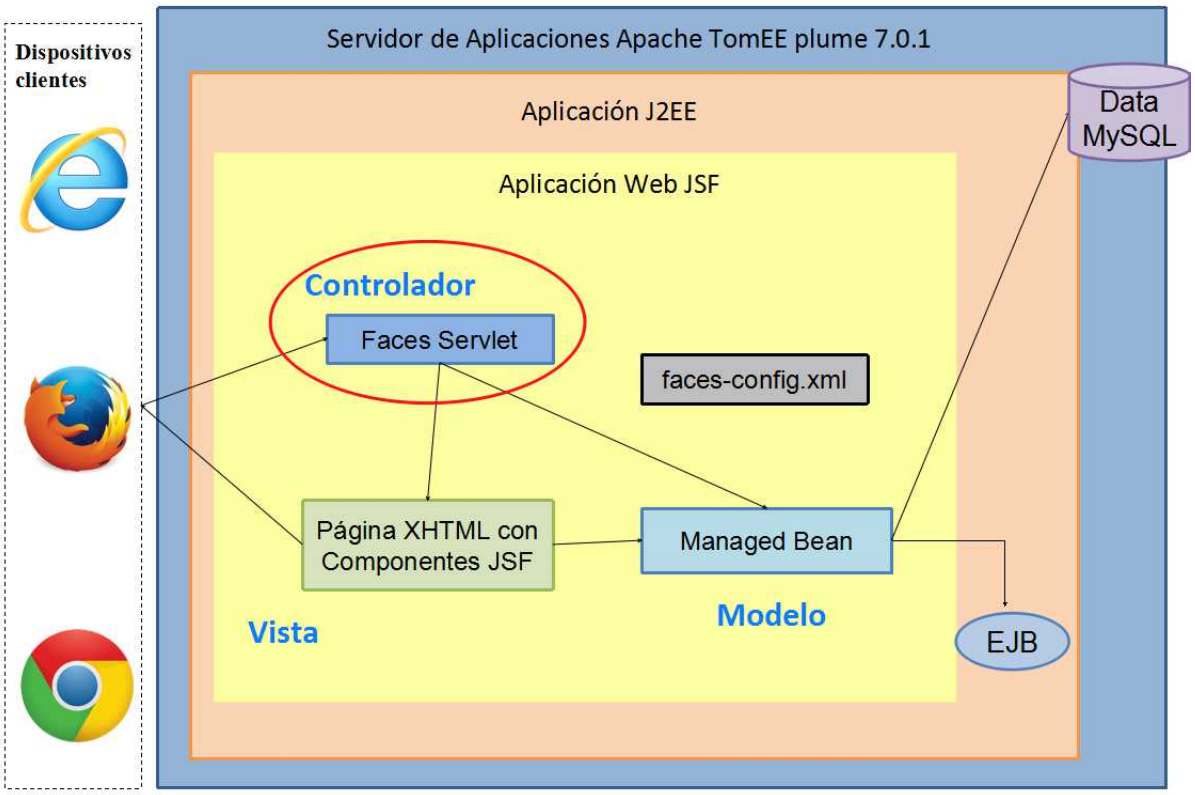


Figura 4.2: *Lógica de negocio dentro del modelo modelo-vista-controlador*

4.3. Comunicación con Back-end y capa de presentación

Para la comunicación con la base de datos se ha utilizado la tecnología JPA. Esta tecnología se comunica con la base de datos a través del protocolo JDBC y su objetivo es facilitar la comunicación de forma que se encapsule la capa de datos para tratarlos como objetos Java. Lo logra creando un mapeo de objetos en función de los atributos definidos para cada clase a las tablas de la base de datos. A estos objetos se les denominan *Plain Old Java Objects (POJOs)*. Define, además de las clases de persistencia para el encapsulamiento de los datos en objetos y la persistencia de los mismos, un lenguaje de consulta: el *Java Persistence Query Language (JPQL)*, para facilitar el acceso a la información. Cada una de las entidades del modelo entidad-relación descrito en la figura 3.13 tiene su clase corres-

pondiente en la codificación de JPA. Estos objetos de la clase son persistentes respecto al sistema de base de datos y se implementan mediante anotaciones o en un fichero .xml para describir el mapeo objeto-relacional. Por comodidad, se han utilizado las anotaciones sobre las clases Java. Todos los atributos deben tener definido el *get* y el *set*, pues es utilizado por el motor de persistencia.

Para la comunicación con la capa de presentación se han utilizado *JavaBeans* donde se encapsulan los objetos necesarios para procesar las peticiones del cliente ("TheoreticalCompoundsController") y enviar las peticiones a la base de datos a través de un *Entrerprise JavaBean* de sesión ("TheoreticalCompoundsFacade"). El ámbito de los *Entrerprise JavaBeans* es de sesión. La capa de presentación se comunica con los *Beans* mediante la tecnología JSF. Se explica dicho intercambio de información en la sección 5.1. Los *JavaBeans* deben cumplir los siguientes requisitos:

- El contructor no puede tener argumentos.
- Solo puede tener atributos de clase privado.
- Todos sus atributos deben ser accesibles mediante métodos *get* y *set* con nomenclatura estándar.
- Debe implementar la interfaz "Serializable".

En la figura 4.3 se aprecian las clases sobre las que se basa la lógica de negocio. El *Bean* utilizado para la búsqueda está definido en la clase "TheoreticalCompoundsController", y se utilizan dos Servlets ("FileUploadServlet" y "DownloadExcelServlet") para el análisis de rutas metabólicas y la descarga del resultado del análisis de las rutas metabólicas a ficheros excel. Esas tres clases son las que implementan la lógica de negocio.

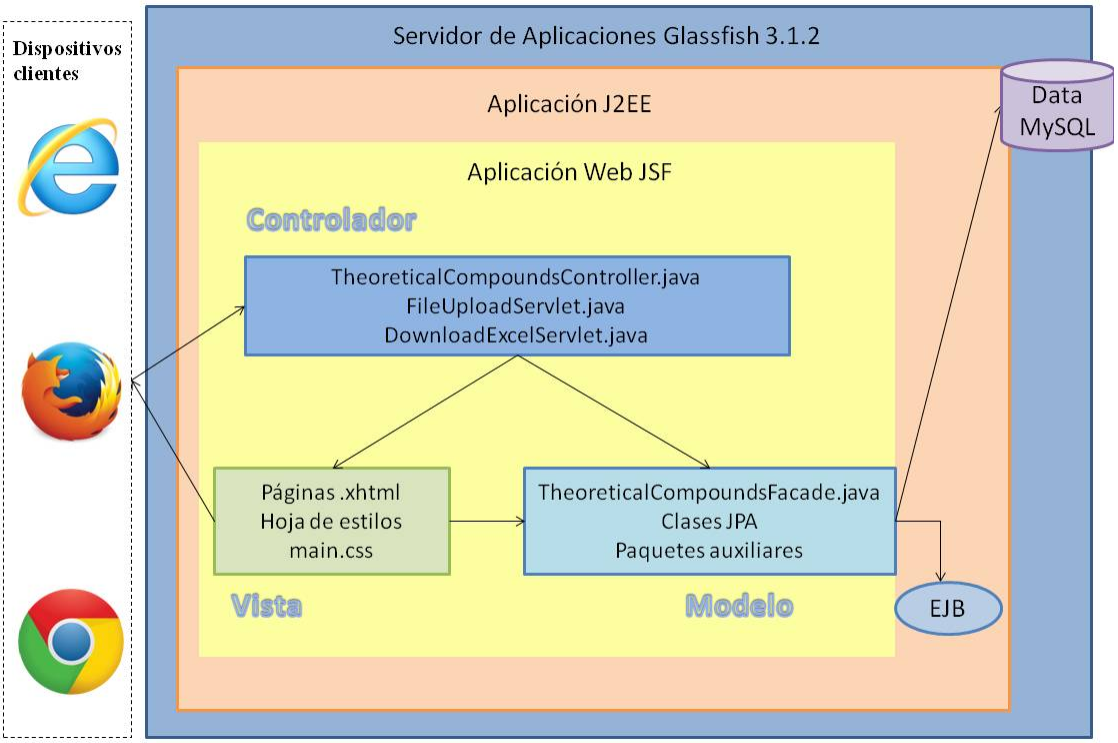


Figura 4.3: *Representación de principales clases Java en el modelo modelo-vista-controlador*

4.4. Back-end

En esta sección se explicarán cada uno de los *sprints* de trabajo correspondientes a la implementación del back-end y la lógica de negocio de la aplicación.

4.4.1. Inclusión de aductos en el motor de búsqueda

Los aductos son alteraciones que se producen en los metabolitos debido a la unión de diferentes moléculas. Las reglas de formación de aductos no son exactas, pero sí hay elementos que es común que se unan a los metabolitos. Aunque las herramientas de Agilent y otras similares facilitadas por las compañías de instrumentación analítica proporcionan algunas funcionalidades para la detección automática de isótopos y aductos y devuelven los resultados de dicho procesamiento en la composición del espectro de cada uno de los picos

generados en el espectrómetro, esta funcionalidad no tiene una fiabilidad absoluta para la detección de isótopos, y la detección de aductos y fragmentos es bastante mejorable.

Para paliar estas deficiencias, es necesario que la herramienta realice una búsqueda de aductos para cada una de las masas detectadas. Esta acción debe estar controlada por la lógica de negocio (ver figura 4.4) Dichos aductos varían en función del producto utilizado para la eletroionización de la muestra. También dependen del modo de ionización (positivo o negativo). Estos elementos están basados en la experiencia previa de los investigadores del CEMBIO, y se corresponde con la lista siguiente para el modo de ionización positivo:

- M+H: hidrógeno.
- M+K: potasio.
- M+Na: sodio.
- M+NH4: amonio.
- M+H-H2O: hidrógeno y pérdida de una molécula de agua (H_2O).
- M+H+HCOONa: hidrógeno y molécula de formiato de sodio (M+H+HCOONa).
- M+2H (Aducto con carga doble): doble átomo de hidrógeno.

Y con los siguientes aductos para el modo de ionización negativo:

- M-H: pérdida de átomo de hidrógeno.
- M+Cl: cloro.
- M+HCOO: anión de formiato (M+HCOO).
- M-H-H2O: pérdida de átomo de hidrógeno y molécula de agua (H_2O).
- M-H+HCOONa: pérdida de hidrógeno y unión con formiato de sodio (M-H+HCOONa).

En la figura 4.4 se puede apreciar como se ha incluido la búsqueda de aductos en función del modo de ionización y los datos que recibe la capa de negocio para la búsqueda simple de aductos. Es necesario, además de las masas experimentales y la tolerancia de dichas masas, el modo de entrada de las masas experimentales, que puede ser neutro, proporción masa/-carga (m/z) o m/z recalculado, así como si el modo de ionización es positivo o negativo y los posibles aductos formados durante el análisis. El m/z recalculado es una de las funciones del software de adquisición de datos, en la cual identifica el modo de ionización y la masa detectada (m/z al ser masa/carga), y calcula las masas neutras sumándole (modo de ionización negativo) o restándole (modo de ionización positivo) un átomo de hidrógeno. La herramienta permite introducir las masas experimentales en cualquier de las tres formas, pues puede haber usuarios que deseen trabajar con masas neutras, con m/z, o con m/z recalculadas.

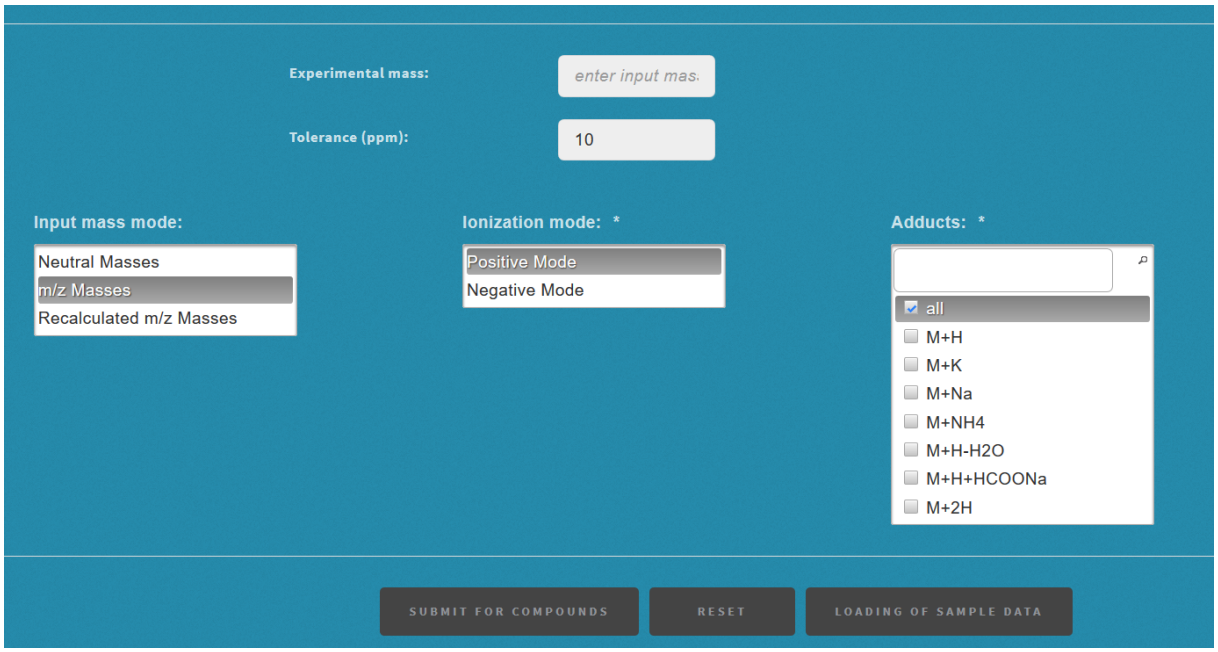


Figura 4.4: *Inclusión de búsqueda de aductos en función del modo de ionización*

Tras conocer qué tipo de masas experimentales ha introducido el usuario, la herramienta debe buscar los posibles metabolitos en función de la masa experimental y de los posibles

aductos. Dichos aductos dependen del experimento, por lo que deben ser seleccionados por el usuario.

La búsqueda de aductos se realiza añadiendo o sustrayendo las masas de los posibles elementos que proceden de la ionización del metabolito. Cuando es posible la formación de aductos debe buscarse la masa del elemento con la adición o sustracción del posible aducto. Se han creado estructuras (mapas), que contienen los posibles aductos y en función del modo de entrada de datos y del modo de ionzación, el Bean donde se implementa la lógica de negocio toma los datos de estas estructuras en función de la elección del usuario.

La búsqueda a la base de datos se realiza a través del EJB de sesión, que tiene un objeto del tipo "EntityManager" para realizar consultas persistentes a los objetos mapeados mediante JPA. Se calcula la tolerancia permitida mediante la fórmula mostrada en 4.1

$$Masa\,a\,buscar = Masa\,experimental \cdot ppm \cdot 10^{-6} \tag{4.1}$$

En caso de que no haya ningún resultado para alguna de las consultas, se crea un objeto "NoCompounds" que indica que no ha habido resultados. Esto es así porque para la generación de ficheros .xls facilita el trabajo tener objetos de la misma clase padre, de la que heredan los diferentes tipos de compuestos, como "NoCompounds". Los compuestos devueltos por la consulta se incluyen en dos listas diferentes, una de ellas es una lista de objetos agrupados (cada grupo corresponde a una masa experimental y un aducto), y la otra una lista que contiene los compuestos. Ambas listas están ordenadas en función de la masa experimental, el aducto, y la diferencia del compuesto respecto a la masa experimental y el aducto correspondiente.

Con esta información, el controlador tiene implementada la búsqueda que debe realizarse sobre las entidades de información y envía a la capa de presentación los resultados de la búsqueda, que pueden apreciarse en las figuras 4.5 y 4.6 y también puede exportar dicha información a formato excel mediante el botón correspondiente en la página de resultados. Los compuestos devueltos son enviados a la capa de presentación mediante una lista de

objetos agrupados que posteriormente procesa la interfaz de presentación como una lista paginada.

Metabolites found for mass 757.5667 and adduct M+NH4 -> 108 metabolites found

Id	Molecular Weight	Flag	error PPM	Cas	Name	Formula	KEGG	HMDB
100902	739.5516	27.755224	25		PC(P-16:0/18:3(9Z,12Z,15Z))	C42H78NO7P		HMDB11213
116737	739.5516	27.755224	25		PC(P-16:0/18:3(6Z,9Z,12Z))	C42H78NO7P		HMDB11212
118762	739.5152	27.755224	24		PE(22:4(7Z,10Z,13Z,16Z)/14:0)	C41H74NO8P		HMDB09580
117306	739.5152	27.755224	24		PE(20:4(8Z,11Z,14Z,17Z)/16:0)	C41H74NO8P		HMDB09418
105663	739.5152	27.755224	24		PE(20:4(5Z,8Z,11Z,14Z)/16:0)	C41H74NO8P		HMDB09385

Figura 4.5: *Resultado de una búsqueda simple en Ceu Mass Mediator (continúa en la figura 4.6)*

HMDB	LipidMaps	Metlin	PubChem	Pathways
HMDB11213	LMGP01030032		52923894	
HMDB11212			53480681	
HMDB09580	LMGP02011104		52924805	
HMDB09418			53479802	
HMDB09385	LMGP02011174		52924875	

Figura 4.6: *Resultado de una búsqueda simple en Ceu Mass Mediator (continuación de la figura 4.5)*

4.4.2. Generación de ficheros .xls

Tras ese primer *sprint* en el que se implementó la búsqueda de aductos, el siguiente objetivo es poder obtener los resultados en ficheros .xls. Es un objetivo prioritario puesto que los químicos analíticos están acostumbrados a trabajar con este tipo de ficheros y eso les permite ahorrar mucho tiempo de su trabajo sin pasar por una curva de aprendizaje previa que supondría el cambio a otro método de trabajo. Para la implementación de este requisito, se ha utilizado la biblioteca POI de apache que proporciona una serie de clases y métodos para trabajar con ficheros cuyo formato es .xls y .xlsx. Para la escritura se generan las dos

primeras líneas estáticas, que contienen el nombre del fichero y el nombre de las cabeceras de cada columna. A partir de ahí se comienza a generar una línea para cada compuesto. Es necesario adaptar la nomenclatura al formato .xls, que utiliza el carácter "," en lugar de "." para los números racionales.

Para la generación de ficheros .xls generados en función del agrupamiento de rutas metabólicas era más útil la biblioteca JExcelApi, ya que permite tratar el fichero excel como una cuadrícula y escribir de forma automática en la columna y celda que se prefiera. La generación de ficheros .xls a partir de rutas metabólicas tiene mayor dinamismo que la generada a partir de búsqueda de compuestos, puesto que respeta el orden de las columnas en las que el usuario ha introducido los datos, siempre y cuando la columna de rutas vaya en la posición última (14 o 15 en caso de que los datos tengan la columna "flag"). Para esta generación dinámica de .xls la biblioteca POI se adaptaba peor que JExcelApi, y se implementó la generación a partir de la agrupación de rutas metabólicas con JExcelAPI. Los resultados se exportan en ficheros con formato .xls debido a que es el formato propio de Microsoft Office para hojas de cálculo para las versiones antiguas a Microsoft Excel 2003, y es también reconocido por las posteriores. Por el contrario, el formato .xlsx o .xlsm es reconocido únicamente por versiones de software posteriores al 2003. El estándar .xls es además leído por programas de licencia gratuita como OppenOffice, XLS reader, XLS Viewer, ...

4.4.3. Integración de nuevas bases de datos

La base de datos HMDB contiene información detallada de 41.933 metabolitos encontrados en el cuerpo humano[41]. Está desarrollada dentro del proyecto *Human Metabolome Project* y sus objetivos son los de identificar y cuantificar nuevos metabolitos desconocidos en el cuerpo humano y verificar la información de los metabolitos ya conocidos. Dicha base de datos proporciona información sobre propiedades químicas de cada metabolito y acerca de la enzima asociada. Además referencia los compuestos a fuentes externas, aunque dicha característica se ha comprobado que no es fiable al tener referencias incorrectas, tal y

como se muestra explica en la sección 2.2. En lo concerniente al interés de integrar dicha base de datos en la herramienta desarrollada, ésta proporciona gran cantidad de identificadores de los metabolitos, datos sobre su taxonomía, los posibles orígenes conocidos donde han sido hallados previamente, la función que tienen en el organismo en que se encuentran o información física de los compuestos, entre otros datos. Para completar la descripción de los metabolitos y facilitar la identificación, para algunos de los compuestos ofrece una predicción de los resultados al aplicar espectrometría en tándem sobre ellos con diferentes técnicas (GC-MS, LC-MS, 1d NMR, 2d NMR) y en condiciones variables del experimento (instrumentración usada, voltaje utilizado, modo de ionización, ...)

HMDB ofrece la posibilidad de descargar sus datos para fines no comerciales en formato .xml. La información de cada metabolito está contenida en un fichero de este formato. El trabajo de la integración consistía en la creación de un analizador sintáctico para documentos .xml. Tras revisar diferentes bibliotecas se optó por la biblioteca DOM de W3C, que permite la lectura de .xml como árboles. Se utiliza el analizador sintáctico desarrollado para extraer la información relativa a cada compuesto e incluirla en la base de datos de CEU Mass Mediator y para la actualización de los compuestos ya existentes en la misma. Se realizó un diseño pensando en la inclusión de conocimiento para aumentar la capacidad de la herramienta en un desarrollo futuro.

4.4.4. Estrategia de unificación de compuestos

El siguiente paso natural en el desarrollo de la aplicación era el tratamiento de los metabolitos como elementos únicos independientemente de la fuente de datos de donde procedían. Esto era así porque la unificación ayudaría en diferentes aspectos: permite reducir el número de entradas en la base de datos; permite reducir el número de compuestos que el usuario de la herramienta debe filtrar una vez obtenga los resultados de su consulta; y, además, permite unificar y completar la información sobre los compuestos existentes en diferentes fuentes. Por ejemplo, el origen del compuesto puede completarse desde HMDB si

no es facilitado por KEGG, o las rutas de los compuestos que no aparecen en LipidMaps pueden completarse con la información desde KEGG. Al fin y al cabo, los compuestos son entidades y las entradas de las diferentes bases de datos son sólo representaciones de los compuestos que facilitan información sobre ellos. Para ello, se hizo un estudio detallado de las posibilidades existentes en la unificación de compuestos con sus ventajas e inconvenientes. A continuación, se va a hacer una descripción cronológica del estudio sobre la unificación.

Existe un identificador del que se ha hablado anteriormente denominado CAS *Registry Number* (**CASRN**) patentado y mantenido por la sociedad Americana de química (**A**merican **C**hemical **S**ociety). La ACS asigna estos número de forma secuencial y en orden ascendente con un formato consistente en tres partes. La primera de ellas tiene de 2 a 7 dígitos, la segunda 2 dígitos, y la última es 1 dígito de control. Estas tres partes están separadas por un guión, y sólo disponen de él aquellos compuestos registrados por la ACS, pero hay numerosos metabolitos, como los lípidos, que no disponen de él. Dicho identificador es único para cada compuesto, pero el acceso a la lista oficial de identificadores CASRN mantenido por la ACS tiene restricciones fuertes en el uso de la información que ellos ofrecen y conlleva un desembolso económico. Aunque desde la Universidad CEU San Pablo se dispone de una licencia para el acceso a esta base de datos, los términos de esa licencia prohíben explícitamente el uso de herramientas automáticas para acceder a los CASRN, así como el almacenamiento local de más de 3000 de estos identificadores. Debido a estos motivos, las bases de datos de metabolómica no tienen acceso directo a la lista de identificadores CASRN, y no detallan cómo toman esa identificación, pero es de suponer que lo hacen en función de información de artículos o a partir de buscadores no oficiales de los que hagan uso. Esto da lugar a incongruencias entre compuestos iguales que proceden de diferentes fuentes y hace que no sea útil a la hora de unificar compuestos de diferentes fuentes, aunque sí puede serlo a la hora de buscar información sobre un determinado metabolito. Por ello, se ha creado una entidad en el modelo de datos llamada compounds_cas para tener la información de los CAS que está disponible en la herramienta facilitada por la ACS. Este conocimiento

puede ser utilizado posteriormente para verificar errores procedentes de las fuentes referente al CAS o a la información del compuesto.

Existe también un código denominado *Smiles* (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification)[63] que nació en 1988 de la necesidad de describir la estructura de las moléculas sin ambigüedades. La idea era la identificación unívoca de compuestos, pero estaba pensada para tener un código a partir del cual se pudiese dibujar la estructura y fuese legible en ASCII. Esa función la cumple a la perfección, pero una misma estructura puede representarse por varios *Smiles* (cuánto mayor sea la molécula, más opciones a la hora de escribir un código smiles para ella), lo que hace que esta especificación no sea útil a la hora de unificar compuestos.

Tras el intento de unificación a partir del CASRN y del código smiles, se procedió a buscar qué otras posibilidades de unificación había, siempre con la condición de que ningún metabolito podía ser unificado si no había una confianza del 100 % en dicha operación, y se intentó a través de los ficheros .mol[59]. Estos ficheros son facilitados por las fuentes y sirven para dibujar los diferentes compuestos. En ellos está contenida la tabla de conexiones, que muestra información sobre el número de átomos, enlaces, el símbolo atómico de cada átomo y los enlaces entre diferentes átomos. Para una completa descripción se recomienda leer la cita sobre dichos ficheros, que es un manual donde está explicado con detalle dicho formato[59]. El problema de los ficheros .mol es que, por convenio, los átomos de hidrógeno pueden estar o no incluidos en la tabla de conexiones del fichero, lo que dificulta la lectura de los ficheros. Los compuestos pueden ser dibujados de infinitas formas, aunque siempre deben tener la misma conectividad, por lo que este segundo problema es subsanable.

Tras hacer un estudio bibliográfico y valorar la opción de la creación de algoritmos para la lectura unificada de este tipo de ficheros se encontró un identificador que era generado a partir de estos ficheros, llamado **IUPAC International Chemical Identifier (InChI)**. Este identificador es generado a partir de diferentes estructuras que existen para la definición de la estructura de una molécula, entre ellas el molfile. Existe un software bajo licencia

LGPL[62] que genera el código InChI a partir de estas estructuras. El principal escollo a superar para poder sacar provecho del InChI es que su longitud varía fuertemente. Este identificador define en diferentes capas la fórmula química, la conexión de los átomos, los átomos de hidrógeno, la carga de los diferentes átomos, la estereoquímica de la molécula y tiene una capa para identificar los isótopos. Las capas del InChI son las siguientes:

1. Capa principal.
2. Capa de carga.
3. Capa de estereoquímica.
4. Capa de isótopos.
5. Capa H mixta.
6. Capa de reconexión.

Las dos últimas capas son opcionales y debe tenerse en cuenta que la versión actual de InChI (versión 1) no soporta los siguientes tipos de compuesto:

- Polímeros.
- Metales orgánicos complejos.
- Estructuras de Markush.
- Mezclas.
- Conformaciones.
- Isómeros con estado excitado.
- Compuestos con estereoquímica o quiralidad no local.
- Isómeros topológicos.

- Organismos polimorfos.
- Conjuntos de moléculas.
- Isótopos con enriquecimiento sin especificar.
- Reacciones.

Para poder trabajar con el InChI Identifier y hacerlo más confortable existe una clave, llamada *InChIKey*, que es el resultado de aplicar un algoritmo hash **SHA-256** sobre la InChI completa. La InChIKey está formada por 14 caracteres que describen la capa de conectividad, un guión, 10 caracteres resultantes de aplicar el algoritmo hash sobre el resto de capas, otro guión, y un último carácter que realiza un chequeo de comprobación (*checksum*). Además, InChI tiene otra ventaja sobre el código CAS o Smiles. Estos dos estándares son propietarios, mientras que InChI es un proyecto abierto y el acceso a sus algoritmos es mucho menos opaco al estar bajo licencia LGPL version 2.1.

Por lo tanto, había una forma de unificar compuestos, pero era necesario que las fuentes de datos facilitasen o bien el InChI del compuesto, o bien el fichero de formato .mol describiendo su estructura a partir del cual se pudiese generar el InChI. En el caso de la herramienta en desarrollo, que incluye cuatros fuentes diferentes, HMDB facilita el InChI y la estructura en formato .mol. LipidMaps proporciona ambas, aunque para acceder al InChI es necesario hacer uso de su API rest, ya que el InChI completo es una información no incluida en los ficheros descargables debido a su excesiva longitud. KEGG no facilita el InChI en la información de cada compuesto, pero sí los ficheros .mol. En cuanto a Metlin, no podíamos acceder a nuevos datos debido a que actualmente su API está fuera de servicio por problemas de seguridad, y se desconoce si volverá a tener una API disponible, así que estos compuestos deben ser tratados como independientes al no disponerse ni poderse generar de ningún modo su InChI.

El resultado de este trabajo fue que la unificación era posible, incluso para compuestos que no tenían CAS, objetivo que no se creía alcanzable en la primera aproximación. En la

figura 3.2 se aprecia como se incluyó en el modelo de datos una entidad para incluir los identificadores (InChI e InChIKey) correspondientes a cada compuesto (entidad llamada compounds). Faltaba entonces implementar un analizador sintáctico para las fuentes que facilitaban el InChI y automatizar el proceso para la generación de las claves a partir del fichero .mol facilitado por KEGG. El software[62] fue utilizado desde código generado en *shell script* para cada uno de los compuestos de KEGG, ya que era la única fuente de datos que facilitaba el fichero .mol de cada compuesto pero no el identificaor InChI. Con el fichero de órdenes (*script*) desarrollado se automatizó la generación del identificador InChI a partir del fichero .mol (Ver apéndice B).

Resultados de la unificación de compuestos

En la tabla 4.1 se aprecia el número de compuestos proveniente de cada fuente de datos y en la figura 4.7 puede apreciarse de forma gráfica la procedencia de los compuestos contenidos en la base de datos local de la herramienta.

Fuente	KEGG	HMDB	LipidMaps	Metlin
Nº Total de compuestos	17637	41514	40719	75447
Nº de compuestos con identificador Inchi	15559	41297	40196	0
Nº de compuestos sin masa	2027	25	2298	0
Porcentaje de compuestos	10.1 %	23.7 %	23.2 %	43 %

Tabla 4.1: *Número de compuestos en base a fuente de datos*

Los compuestos que no tienen un identificador corresponden a elementos cuyo tipo pertenece a los listados en 4.4.4, o a compuestos cuya fuente no proporciona la estructura de los mismos. Esta segunda opción ocurre en algunos compuestos genéricos o por falta de información en la fuente, como pasa en el caso de Metlin, que no facilita el identificador InChI ni la estructura de los compuestos. No hay posibilidad de unificar un compuesto sin estructura si no es de forma manual o confiando en las referencias a otras bases de datos que proporcione la fuente, referencias que, lamentablemente, se ha comprobado que no son fiables, ya que contiene bastantes fallos (ver sección 2.2). Esta integración manual de información es algo

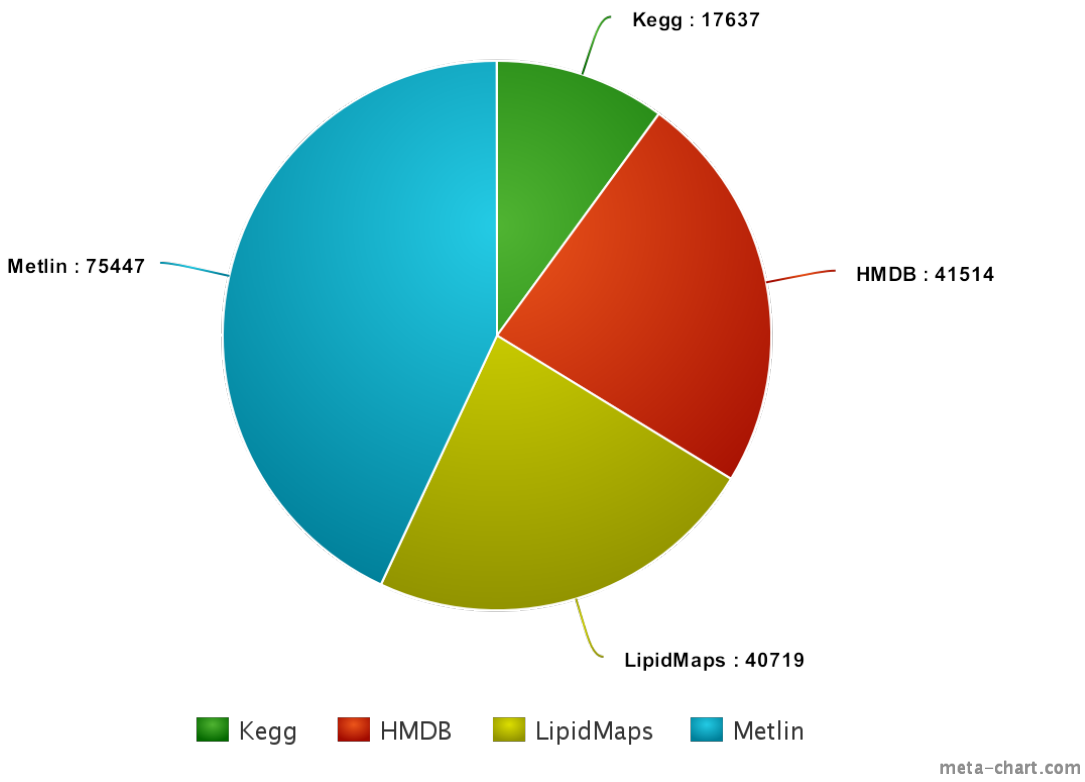


Figura 4.7: *Compuestos totales en la base de datos*

que se estudiará en la sección dedicada a las líneas futuras.

Para efectuar el análisis de los compuestos unificados se van a dejar fuera las estadísticas de los compuestos de Metlin, puesto que, como se ha explicado en el párrafo anterior, no tienen estructura y no son unificables, y se va a tomar como referencia los compuestos que tienen identificador InChI y que tienen masa. Sólo hay 29 compuestos de LipidMaps, 8 compuestos de HMDB y 0 compuestos de KEGG que tienen masa y no tienen identificador. Esto es debido a que por lo habitual los compuestos que no tienen identificador son compuestos genéricos que describen posibles formaciones y, por tanto, su masa es variable y en la base de datos no se especifica dicha masa. Al ser una muestra menor al 0.1 % el número de estos compuestos en el peor de los casos (LipidMaps), no es algo que tenga significancia a la hora de analizar nuestros resultados. En la tabla 4.2 y en el dia-

grama de Venn de la figura 4.8 se puede apreciar el número de compuestos unificados. En la tabla 4.3 se aprecia en porcentaje el número de compuestos unificados. Este porcentaje está calculado en base al número máximo de compuestos unificables (15559 cuando en la relación están los compuestos de KEGG, 40196 cuando están compuestos de LipidMaps y no de KEGG y 41297 para compuestos repetidos en HMDB) según la relación a analizar (KEGG-KEGG, HMDB-HMDB, LipidMaps-LipidMaps, KEGG-HMDB, KEGG-LipidMaps, LipidMaps-HMDB y KEGG-HMDB,LipidMaps). En el caso de la unificación entre las tres fuentes de datos se toma como máximo número de unificaciones el número de compuestos de KEGG, al ser el menor y marcar el límite máximo. Aunque a primera vista el número de compuestos unificados pueda parecer pequeño (19.83 % entre KEGG y HMDB, 12.91 % entre KEGG y LipidMaps, 0.5 % entre HMDB y LipidMaps y un 4.41 % entre los tres), es importante hacer notar que esta unificación ha tomado en cuenta la estructura completa del compuesto. Esto hace que, si hay pequeñas variaciones en el número de dobles enlaces, en la posición de los mismos o en la capa de reconexión, tome los compuestos como diferentes, ya que realmente lo son. En las referencias facilitadas en las fuentes hacia otras bases de datos, muchas veces esto no ocurre, y, para obtener información de un mismo compuesto en varias fuentes de datos, los químicos deben realizar búsquedas en diferentes fuentes y comprobar la estructura manualmente. Este nivel de especificidad a la hora de unificar es útil para los químicos, pues podrán obtener información de varias fuentes sabiendo que realmente provienen del mismo compuesto a través de una misma página, evitando también posibles errores humanos al identificar estructuras.

Por otra parte, es muy importante que no exista la posibilidad de unificación sin una seguridad completa y es el proceso posterior a la búsqueda dónde se filtran los resultados en función del conocimiento experto y técnicas como el análisis de fragmentación en tándem. Es en esta etapa posterior a la búsqueda en bases de datos donde deben descartarse los compuestos que no han sido unificados en el proceso automático al no tener la misma estructura y que, debido a sus propiedades, no se corresponden con la masa experimental.

	KEGG	HMDB	LipidMaps	TODAS
KEGG	85	3085	2008	686
HMDB	3085	207	4049	686
LipidMaps	2008	4049	56	686

Tabla 4.2: *Relación de compuestos unificados según fuente de datos*

	KEGG	HMDB	LipidMaps	TODAS
KEGG	0.54 %	19.83 %	12.91 %	4.41 %
HMDB	19.83 %	0.5 %	10.07 %	4.41 %
LipidMaps	12.91 %	10.07 %	0.14 %	4.41 %

Tabla 4.3: *Relación de compuestos unificados según fuente de datos*

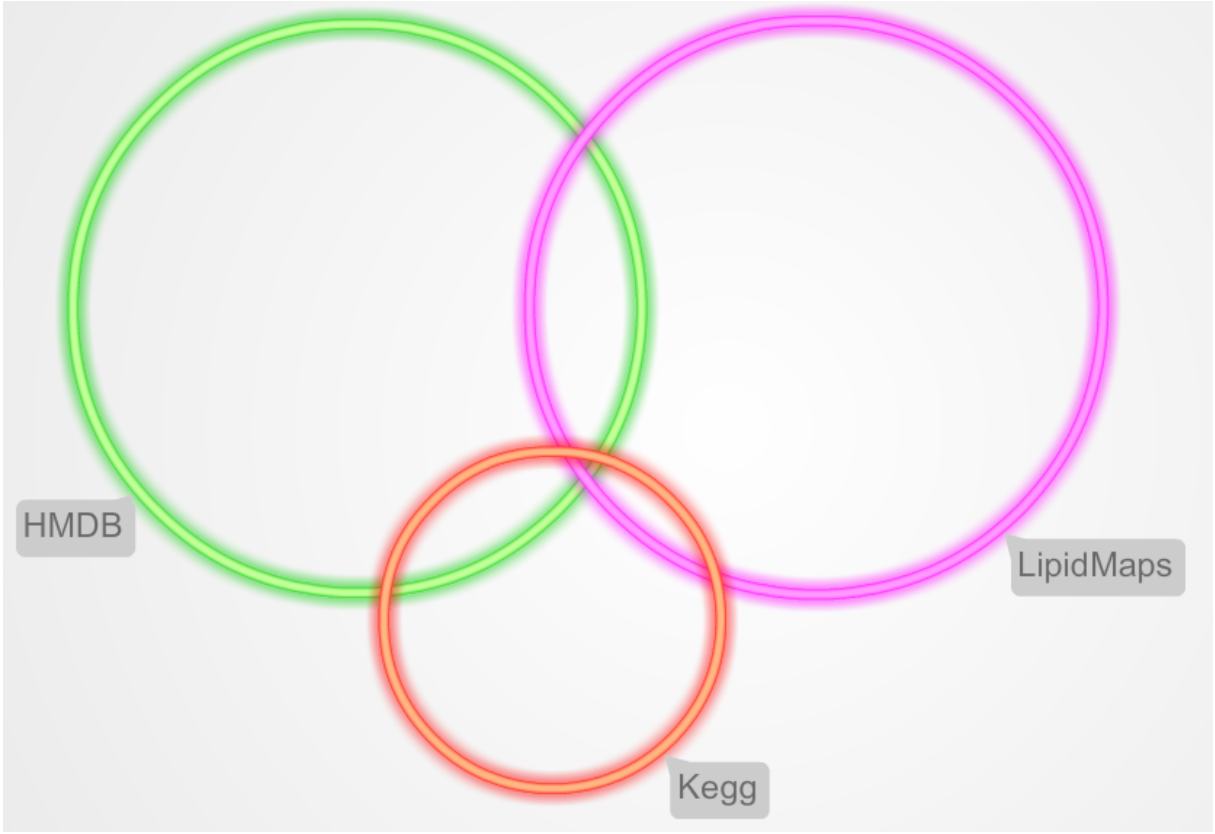


Figura 4.8: *Compuestos totales en la base de datos*

4.4.5. Búsqueda avanzada

Se utilizan las mismas tecnologías que en la búsqueda básica, pero se incluye información sobre tiempos de retención, espectro de composición para cada metabolito detectado y el alfabeto químico que forma la muestra a analizar. En la figura 4.9 se puede ver la información que se envía en la petición sobre búsqueda avanzada de metabolitos. Por analogía, la muestra de resultados es igual a la mostrada para la búsqueda simple de metabolitos, añadiendo una columna ("Flag") para una clasificación personalizada para cada metabolito (usualmente el tiempo de retención).

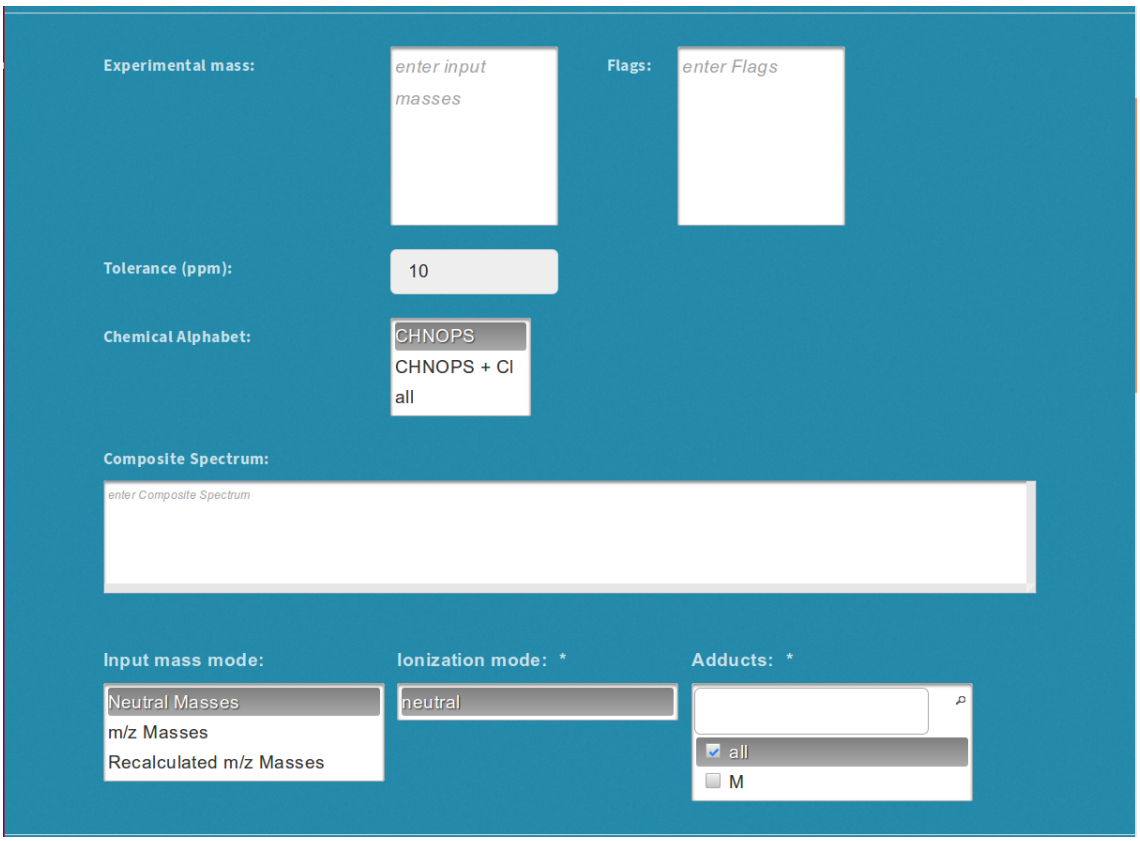


Figura 4.9: *Presentación de la búsqueda avanzada de metabolitos*

Se hace uso de la composición del espectro para la detección de los aductos con doble carga detectados por las herramientas software de adquisición de datos. Algunas de estas completan la composición del espectro y detectan la proporción m/z cuando la entrada de

datos es m/z recalculada. En ese caso los aductos con doble carga se detectan automáticamente y se recalcula su masa.

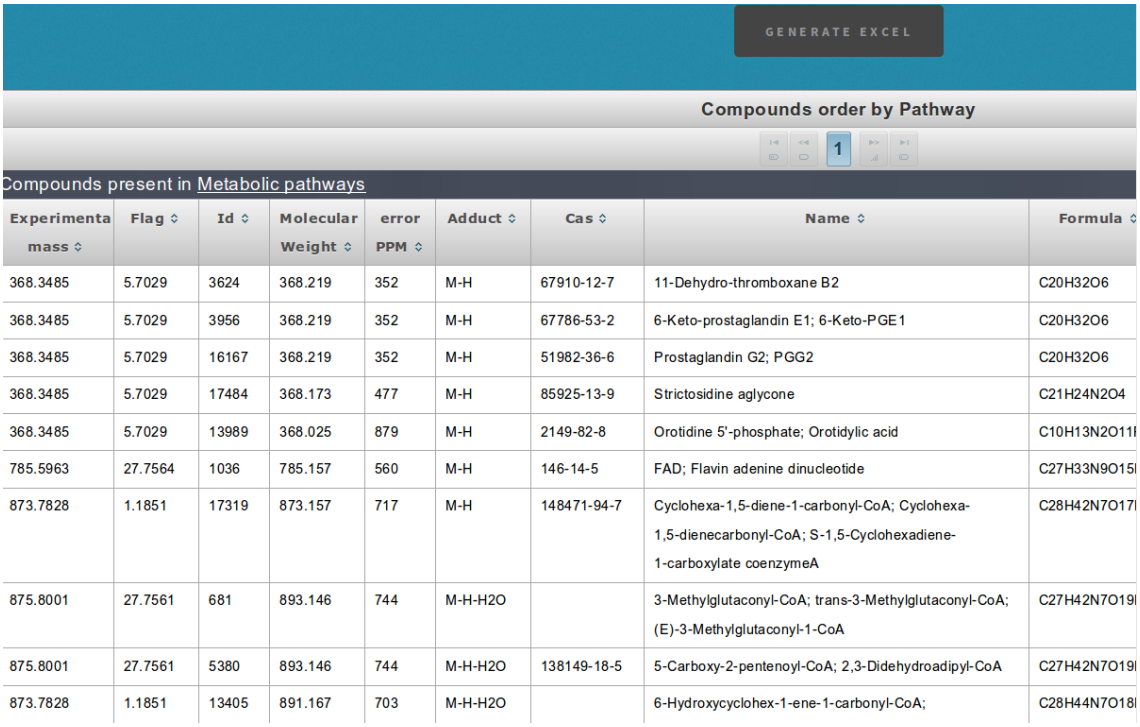
El alfabeto químico es utilizado para filtrar la búsqueda de compuestos y limitarla únicamente a compuestos que pertenezcan a cada uno de los grupos disponibles (CHNOPS, CHNOPS+Cl o todos los elementos). Esto se realiza mediante expresiones regulares a la hora de la inserción de los compuestos en la base de datos. Mediante estas expresiones se buscan los elementos disponibles en una lista de elementos creada y guardada en una estructura constante para así determinar qué tipo de elementos hay en cada compuesto. También detecta si en la fórmula del compuesto se encuentran elementos que no existen y se guarda la información en un registro de sucesos (*log*). Habitualmente esto ocurre con compuestos genéricos como las fosfocolinas, las cuales tienen estructuras unidas a una colina que puede estar definida como genérica en las bases de datos. Estos compuestos no tienen masa, puesto que tienen elementos no definidos (con peso desconocido). Aunque se guardan en la base de datos porque pueden ser útiles en el futuro, al no tener masa, no influyen en las consultas de bases de datos, puesto que se realizan a partir del campo de la masa, campo indexado, minimizando así el impacto en el rendimiento de la aplicación. Dichas expresiones regulares están implementadas mediante la biblioteca "java.util.regex". Se procesan los elementos una única vez al insertar para la clasificación de los compuestos y para obtener luego elementos de determinado tipo se hace una restricción en la consulta sobre el tipo deseado.

4.4.6. Análisis de rutas metabólicas

Desde el comienzo del estudio de la metabolómica el principal objetivo es conocer qué elementos están presentes en las muestras a analizar y poder establecer relaciones entre estos elementos para probar o descartar la hipótesis de partida. A estas relaciones entre metabolitos se las conoce como rutas metabólicas o *pathways*, ya explicadas en el capítulo 1. El análisis de rutas metabólicas trata de establecer las conexiones que hay entre los metabolitos detectados para poder formar un mapa completo de las distintas rutas biológicas y el impac-

to de las rutas sobre la hipótesis inicial. En las herramientas disponibles para el análisis de rutas metabólicas se procesan los datos con diferentes algoritmos para evaluar la posibilidad de que el origen del elemento sea dicha ruta y evaluar por otra parte la importancia de la ruta en el análisis completo. En herramientas como Agilent Mass Profiler Professional los algoritmos son privados y no están accesibles. Pero hay otras herramientas como metPA, cuyos métodos si son públicos[71], GenMAPP[49] o Pathway Commons[8], que permiten evaluar el papel de los elementos procedentes de los grupos de muestras con respecto a la ruta donde puede aparecer y la importancia de la ruta en función de los elementos presentes dicha ruta que al mismo tiempo forman parte del conjunto de las muestras analizadas. Para ello aplican métodos estadísticos como el test exacto de Fisher, el test hipergonómico, el test global o la Ancova global. Tras estos tests, la herramienta aplica un análisis univariante a nivel de compuestos (T test, análisis univariante y regresión lineal) para facilitar una vista detallada de la distribución de los metabolitos y su concentración. Estas herramientas no están siendo usadas actualmente por el cliente por falta de fiabilidad en los resultados, por lo que los análisis se realizan de forma manual. Por ello, en este proyecto se ha desarollado un análisis de rutas metabólicas sencillo que consiste en una agrupación de compuestos en función de la ruta metabólica y una ordenación de estos en base al número total de elementos que aparecen en la muestra. Este análisis da a los químicos analíticos una idea sobre el orden en el que deben buscar las posibles rutas y, en consecuencia, reacciones, que han tenido lugar en el organismo humano.

La lógica de negocio referente al análisis tiene como entrada un fichero excel, que es la información enviada en la petición al *Servlet* con nombre "FileUploadServlet". Este fichero debe tener el formato definido en la tabla 3.6, y devuelve a la capa de presentación una respuesta que redirecciona la petición a la muestra del resultado del análisis de rutas metabolómicas, como se aprecia en la figura 4.10. Es posible descargar los resultados de la agrupación por rutas metabólicas a un fichero excel, petición que procesará otro Servlet con nombre "DownloadExcelServlet".



Experimental mass	Flag	Id	Molecular Weight	error PPM	Adduct	Cas	Name	Formula
368.3485	5.7029	3624	368.219	352	M-H	67910-12-7	11-Dehydro-thromboxane B2	C20H32O6
368.3485	5.7029	3956	368.219	352	M-H	67786-53-2	6-Keto-prostaglandin E1; 6-Keto-PGE1	C20H32O6
368.3485	5.7029	16167	368.219	352	M-H	51982-36-6	Prostaglandin G2; PGG2	C20H32O6
368.3485	5.7029	17484	368.173	477	M-H	85925-13-9	Strictosidine aglycone	C21H24N2O4
368.3485	5.7029	13989	368.025	879	M-H	2149-82-8	Orotidine 5'-phosphate; Orotidylic acid	C10H13N2O11
785.5963	27.7564	1036	785.157	560	M-H	146-14-5	FAD; Flavin adenine dinucleotide	C27H33N9O15
873.7828	1.1851	17319	873.157	717	M-H	148471-94-7	Cyclohexa-1,5-diene-1-carbonyl-CoA; Cyclohexa-1,5-dienecarbonyl-CoA; S-1,5-Cyclohexadiene-1-carboxylate coenzymeA	C28H42N7O17
875.8001	27.7561	681	893.146	744	M-H-H2O		3-Methylglutaconyl-CoA; trans-3-Methylglutaconyl-CoA; (E)-3-Methylglutaconyl-1-CoA	C27H42N7O19
875.8001	27.7561	5380	893.146	744	M-H-H2O	138149-18-5	5-Carboxy-2-pentenoyl-CoA; 2,3-Didehydroadipyl-CoA	C27H42N7O19
873.7828	1.1851	13405	891.167	703	M-H-H2O		6-Hydroxycyclohex-1-ene-1-carbonyl-CoA;	C28H44N7O18

Figura 4.10: *Resultado del análisis de metabólicas*

El back-end se limita a hacer uso del modelo de datos que puede verse en la figura 3.13 para agrupar los compuestos en función de las rutas metabólicas donde están categorizados o identificados. Esta agrupación tiene lugar en la lógica de negocio. Está agrupación de compuestos es enviada luego a la capa de presentación para mostrar los resultados y puede ser descargada como fichero .xls.

4.4.7. Automatización del refresco de datos y de las copias de seguridad

Al finalizar la implementación del back-end, disponer del modelo de datos implementado y la información incluida en nuestra base de datos, era conveniente automatizar todo el proceso de extracción de datos desde las fuentes originales, transformación de los datos originales a nuestro modelo de datos, integración de los metabolitos y carga en nuestra base de datos para que pudiese realizarse en un futuro con la menor intervención de personal,

con el consecuente ahorro de recursos temporales y humanos. Se han creado varios códigos de *shell script* para la descarga automática de los recursos remotos y la ejecución del código Java correspondiente a la población de los datos recogidos desde las fuentes. Esto incluye también el chequeo de los identificadores InChI y la generación de los mismos por si los compuestos han variado en la base de datos de origen. La generación de los identificadores es completamente automática para los elementos que no tienen esta información en la fuente, como los compuestos de KEGG o los de LipidMaps. También se generan ficheros de registro de eventos no esperados para facilitar la depuración ante posibles problemas durante su ejecución.

En cuanto a la política de seguridad, se ha generado un Crontab en el servidor de producción que realiza una copia de seguridad semanal tanto del proyecto desplegado en el servidor de aplicaciones como de los datos que hay en la base de datos. Los ficheros de seguridad se envían mediante el protocolo scp al servidor de desarrollo, que hace la función de servicio de backup remoto. El crontab está definido de la siguiente forma:

```
0 0 * * 0 $DIR/bk_mediator.sh > $DIR/logs/last_crontab.log
```

Capítulo 5

Lógica de Presentación

La lógica de presentación es la capa del modelo modelo-vista-controlador que hace referencia a la vista que tiene el usuario de la aplicación. Se explica en este capítulo cómo se presentan al usuario los objetos controlados por la lógica de negocio relatada y procesados en el back-end, ambos procesos detallados en el capítulo 4. En la sección 3.2 se detallaron los *wireframes* realizados para el diseño de la herramienta y es en este prototipo en el que se ha basado la lógica de presentación, pues el objetivo a cumplir lo determinaba el resultado del diseño de la interfaz de usuario. Las elecciones de la tecnología se describen en la sección 3.3.1. El objetivo de la lógica de presentación es crear una interfaz fácil de usar por parte del usuario y cuya comunicación con la lógica de negocio sea consistente y no propensa a errores. Los errores de comunicación quedarán registrados en el registro de sucesos (*log*) del servidor de aplicaciones. También se recogen las sugerencias y fallos detectados por el usuario en una cuenta de correo electrónico indicada en la cabecera de la página.

5.1. JSF

JavaServerFaces es una tecnología para aplicaciones Java que simplifica el desarrollo de la capa de presentación de las aplicaciones Java EE. Gestiona su ciclo de vida en el servidor, donde tiene todos sus componentes de control. Proporciona un conjunto de APIs para la representación de componentes predefinidos (formularios, botones, tablas y listas de datos, ...) y para administrar eventos o estados de los que se han hecho uso para la generación de

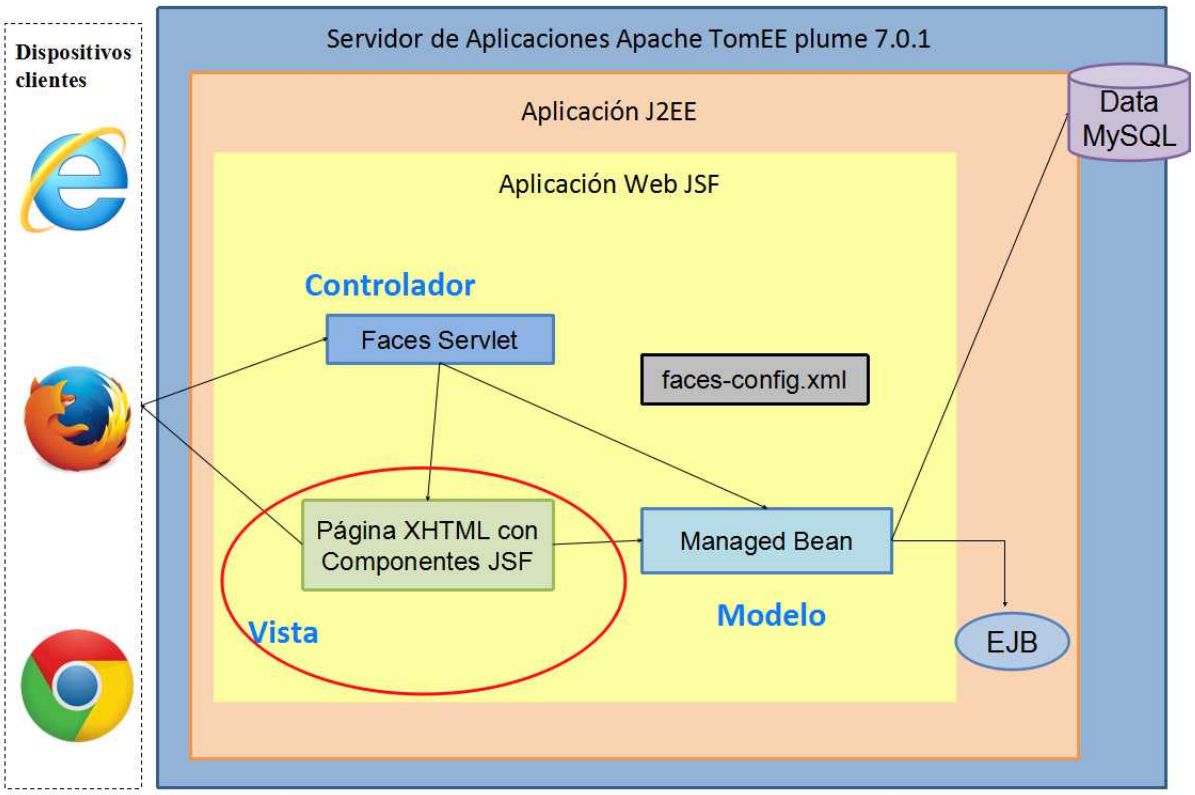


Figura 5.1: *Lógica de presentación en el modelo modelo-vista-controlador*

la interfaz. Se han definido eventos para la validación de datos enviados en los formularios. En este trabajo, todas estas acciones se han servido de la biblioteca de JSF PrimeFaces.

Se utiliza un único Bean ("TheoreticalCompoundsController.java") para la lógica de negocio y dos Servlets. Para la comunicación con ellos se utilizan componentes JSF desplegados en páginas XHTML. Estos componentes JSF hacen referencia a las bibliotecas de etiquetas, que proveen código ya desarrollado para simplificar la presentación de las páginas. Las bibliotecas de etiquetas incluidas en el desarrollo de la lógica de presentación están definidos en la cabecera HTML de las páginas:

```
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:ui="http://java.sun.com/jsf/facelets"
      xmlns:h="http://java.sun.com/jsf/html"
      xmlns:f="http://java.sun.com/jsf/core"
      xmlns:p="http://primefaces.org/ui"
      xmlns:pt="http://xmlns.jcp.org/jsf/passthrough">
```

Se utiliza el primer espacio de nombres (*namespace*) para los componentes de XHTML; el segundo para los componentes de facelets como es la composición de páginas para realizar páginas con componentes reutilizables; el tercero para componentes básicos de HTML como formularios; el cuarto para componentes AJAX para añadir dinamismo a la página y que se puedan modificar elementos por medio de funciones en la misma; el quinto para la biblioteca que facilita PrimeFaces para la utilización de sus componentes previamente desarrollados; y el último para añadir funcionalidades de HTML5 a la página, concretamente a los elementos de PrimeFaces. Los envíos y peticiones de información al *Bean* se hacen a través de las etiquetas previamente explicadas. Todas ellas tienen implementadas funciones para la petición y envío de información al modelo. Son capas independientes y el diseño de la presentación es completamente independiente a la lógica de negocio, siguiendo las normas del modelo modelo-vista-controlador. La estructura de las páginas XHTML corresponde con el menú mostrado en la figura 5.2 y las diferentes funcionalidades que ofrece la herramienta para listar compuestos a partir de sus masas atómicas y agruparlos en función de su ruta metabólica.

Figura 5.2: *Menú principal de la aplicación*

5.2. Front-end

El Front-end se define como la parte del software con la que el usuario interactúa. En caso de una herramienta web, corresponde con la visión del cliente que el usuario tiene de

dicha herramienta desde el navegador.

Para el desarrollo de la interfaz web se utilizó una plantilla proveniente de HTML5UP[18] que tiene una licencia Creative Commons Attribution 3.0 License que permite su utilización libre siempre y cuando se referencie. Esta plantilla estaba construida bajo HTML5 y CSS3 y hacía uso de preprocesamiento mediante directivas SASS, una extensión de CSS para poder usar variables, estructuras de control de flujo y otras funcionalidades. Para su uso es necesario tener instalado Ruby. Al no proporcionar funcionalidades relevantes para el desarrollo de la herramienta, se han usado las directivas SASS en la plantilla, pero no se han añadido nuevas directivas. Contenía un modelo adaptativo (*responsive*) para acomodarse a las dimensiones de dispositivos móviles y tabletas (*tablets*), pero no estaba configurada para el uso de XHTML. Al estar el autor más acostumbrado al desarrollo con estándares estrictos y querer realizar un desarrollo de la interfaz usando diferente composiciones, se adaptó dicho modelo al estándar XHTML.

Los datos que recibe la interfaz de presentación son recogidos por los componentes XHTML definidos mediante etiquetas. Estas etiquetas permiten acceder al Bean "TheoreticalCompoundsController" y sus atributos mediante sus métodos *get* y *set* y actualizar la página dinámicamente. Éstos son tratados como cadenas de caracteres para la visualización, aunque en la capa de la lógica de negocio los tipos de datos pueden ser primitivo u otro tipo de objeto definido por el usuario. En caso de no utilizar tipos primitivos, los objetos definidos por el usuario deben ser procesados según el desarrollo que se establezca por el usuario. En el caso de esta herramienta se utilizan listas y tablas definidas por la biblioteca PrimeFaces para presentar los datos que están guardados como listas de compuestos o listas agrupadas de compuestos. La biblioteca AJAX tiene funcionalidades implementadas que se usan para un recargo dinámico de la página en función de la interacción con el usuario. Se ha hecho uso de elementos de HTML5, especialmente para la inicialización de componentes de PrimeFaces. Estas funcionalidades han dado lugar a un modelo adaptativo que hubo que modificar para la muestra de resultados, pues la información representada en tablas debe

tener una fuente con un tamaño legible y aporta una mayor usabilidad el moverse en la pantalla mediante una barra de desplazamiento (*scroll*) que el dividir las filas de las tablas y, en consecuencia, los datos de cada compuesto. Esto es así para todos los dispositivos y se configura en la hoja de estilos para estandarizar la vista de las diferentes páginas.

Para los dispositivos móviles y tabletas los campos y botones se muestran en diferentes filas en lugar de juntos para que la vista no se sobrecargue y se pueda crear una sensación de sobreinformación al usuario. Para ello existen directivas definidas en la hoja de estilos ("main.css") que se aplican en función del tamaño de la pantalla. Este tamaño de la pantalla se divide en 12 divisiones verticales que pueden utilizarse atributos de estilo para indicar el número de divisiones a utilizar y que las proporciones del diseño se mantengan. Los atributos utilizados tienen un número del 1 al 12 como prefijo y una "u" para indicar el número de columnas a utilizar. Esto hace que el tamaño de los componentes se adapte al tamaño de la pantalla y, además, el tamaño de la pantalla hace que el estilo y colocación de los componentes también sea dependiente del tamaño de la pantalla, generando un efecto dinámico en la interfaz de presentación. Con el uso de identificadores sobre etiquetas para clasificar de forma grupal o individual componentes se han aplicado estilos distintos. Se ha especificado el cambio de estilo de componentes en función del tipo de dispositivo cliente, basándose para la identificación del dispositivo en el tamaño de la pantalla.

El resultado del back-end puede verse en las figuras 5.3, 5.4, 5.5 y 5.6, que perenecen a imágenes extraídas de la página Ceu Mass Mediator (http://ceumass.eps.uspceu.es/mediator/). Las figuras 5.3, 5.4 y 5.5 muestran las diferentes vistas en función del dispositivo desde el que se está accediendo a la aplicación. En la figura 5.6 se aprecia el resultado de la búsqueda de determinadas masas experimentales. Se puede apreciar como esta vista tiene un *scroll* horizontal. Esto es debido a que la información sobre cada metabolito es muy amplia y la mejor solución para mostrar tal cantidad de información es el desplazamiento lateral por la pantalla dentro de una tabla. Hay una tabla diferente por cada posible aducto formado y una lista paginada con las diferentes masas.

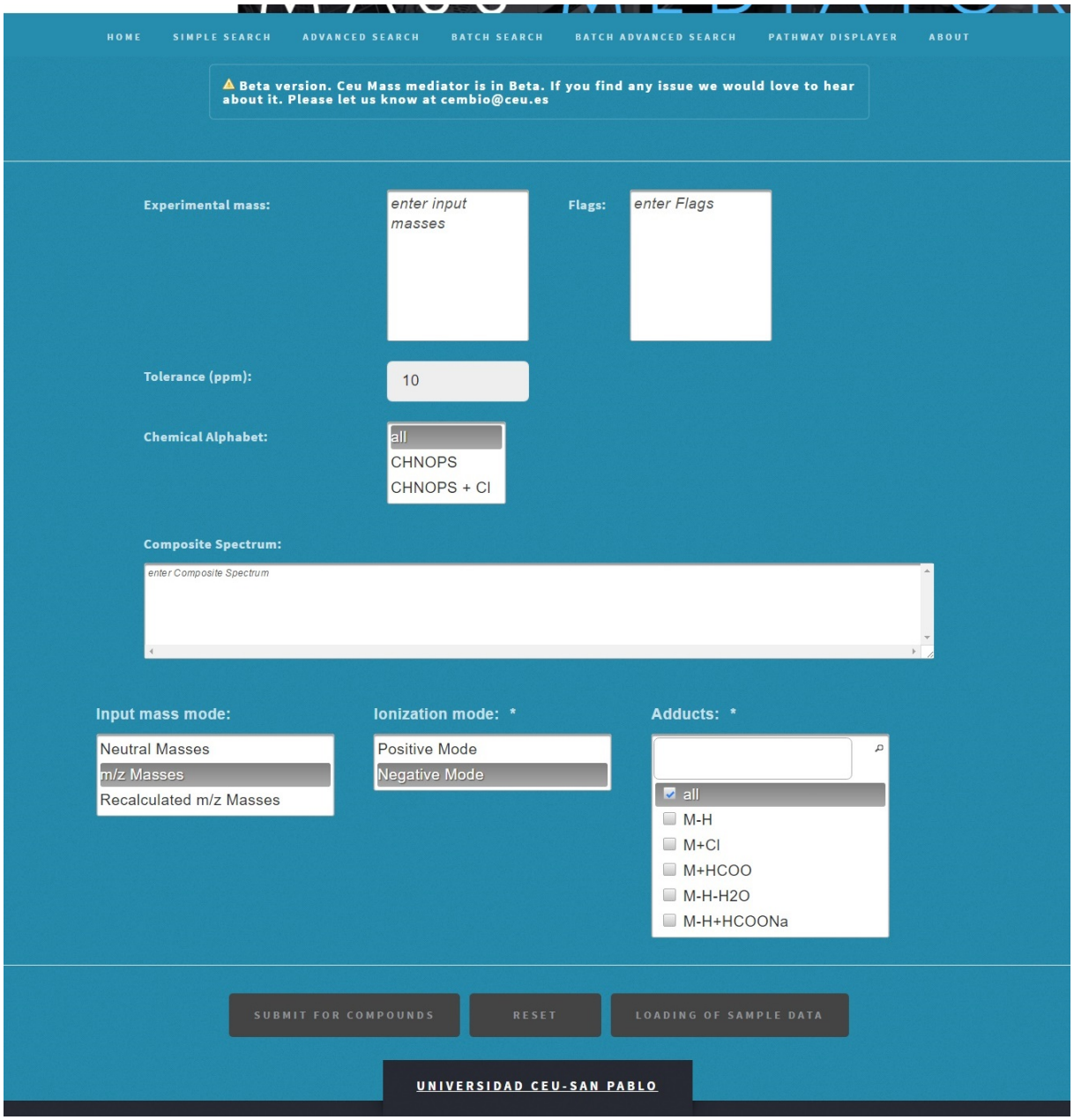


Figura 5.3: *Vista de la aplicación en un dispositivo de tipo escritorio*

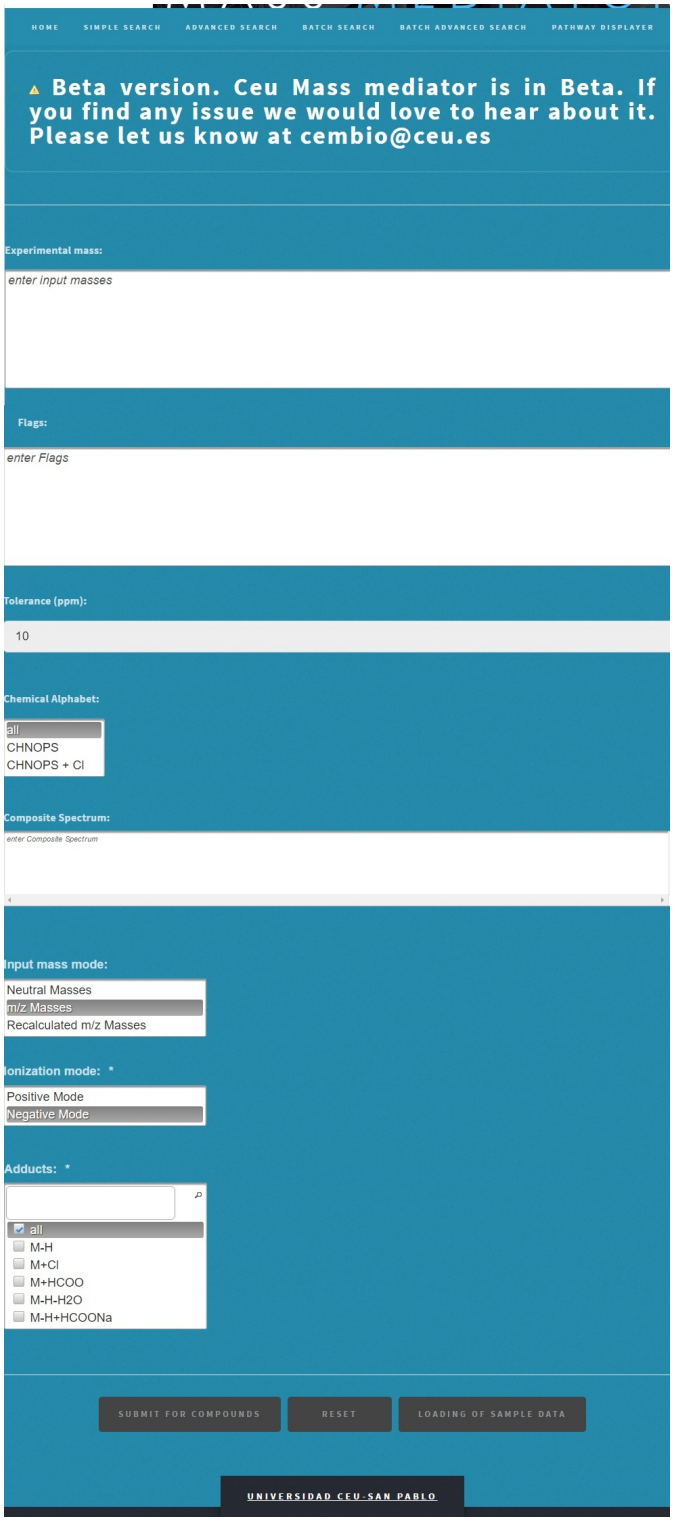


Figura 5.4: *Vista de la aplicación en un dispositivo de tipo tableta*

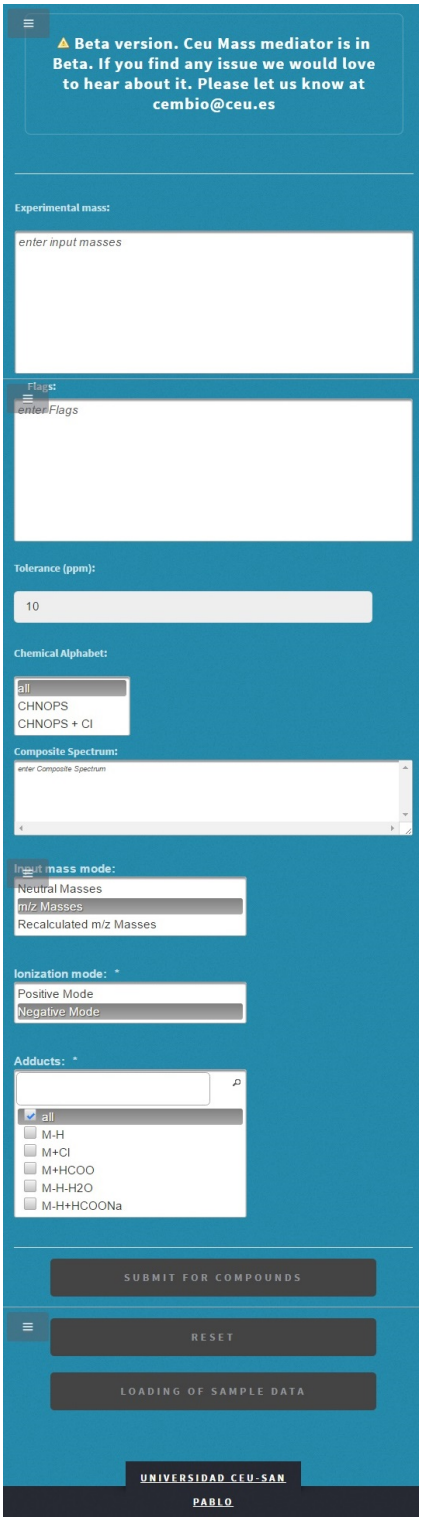


Figura 5.5: *Vista de la aplicación en un dispositivo de tipo móvil*

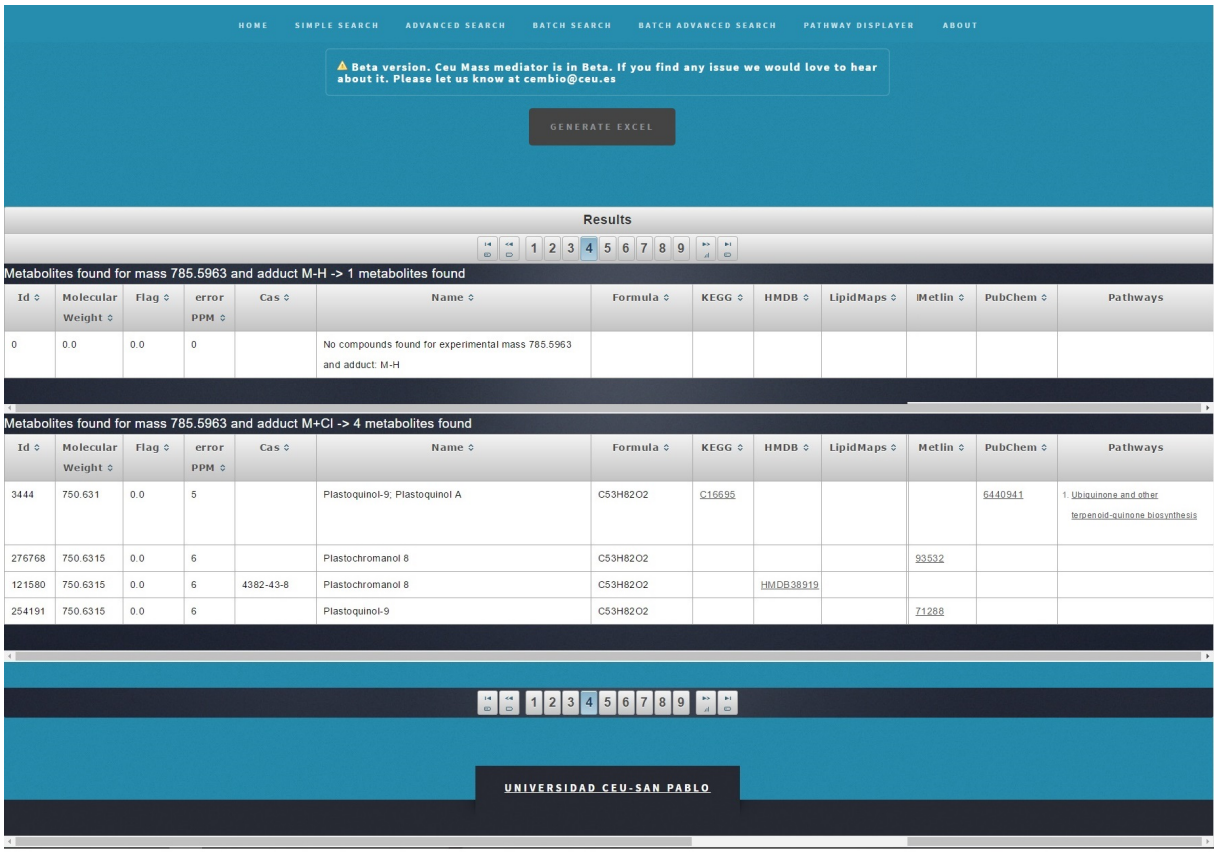


Figura 5.6: *Página de resultados*

Capítulo 6

Resultados y líneas futuras

En este capítulo se va a hacer un análisis de los resultados obtenidos, y se hará un repaso de las posibles líneas futuras para continuar el proyecto presentado en esta memoria.

6.1. Resultados

El objetivo del presente proyecto era el desarrollo de una herramienta web que simplificase y automatizase la búsqueda e identificación de metabolitos, sirviendo así de apoyo y ahorrando una gran cantidad de tiempo al químico analítico. Para ello se ha construido una herramienta que permite realizar búsquedas de modo automático a partir de los datos proporcionados por el espectrómetro de masas sobre múltiples bases de datos de metabolómica.

El proyecto partió de una aplicación básica previa explicada en la sección 1.2, pero debido a la antigüedad de la misma ha sido necesaria cambiarla casi por completo, pues los elementos tecnológicos que utilizaba estaban en la mayoría de los casos obsoletos. Esta aplicación en la actualidad prácticamente no contiene elementos de su versión anterior, bien por la razón previamente explicada sobre tecnologías obsoletas, bien porque era necesario implementar nuevas funcionalidades para las cuales el proyecto anterior no era la opción más conveniente. Durante el desarrollo de este proyecto de fin de máster se han realizado las siguientes tareas:

1. Búsqueda y revisión bibliográfica mediante las herramientas EBSCO Discovery y Scho-

lar Google. Para la comprensión de las necesidades del laboratorio ha sido necesaria la lectura de artículos científicos donde se explican la situación actual de la metabolómica, así como elementos más básicos sobre la química necesarios para poder comprender, analizar y aportar soluciones al campo específico del proyecto.

2. Creación de nuevo modelo de datos para el soporte de la integración de metabolitos. Para ello se ha modificado todo el back-end de la aplicación previamente disponible. Este nuevo modelo facilita la integración de nuevas fuentes de datos, ya que todos los compuestos son tratados como una misma entidad, independientemente de su procedencia. También permite ampliar el conocimiento sobre cada compuesto facilitado por las distintas fuentes, algo que anteriormente tenía que realizarse manualmente por parte de los químicos analíticos.

3. Se ha realizado una integración de las siguientes bases de datos de metabolitos:

   - KEGG.
   - Metlin.
   - LipidMaps.
   - HMDB.

   Esta integración de los datos requiere de una unificación de los compuestos en un único modelo de datos.

4. Unificación en único modelo de datos de compuestos procedentes de dichas bases de datos. Unificación basada en el InChI. Esta unificación reduce los errores humanos y se basa en la estructura de los compuestos. El InChI es un código creado para ser tratado computacionalmente y se toma ventaja de ello para la unificación de compuestos de distintas fuentes.

5. Renovación de la interfaz web de la herramienta utilizada hasta el momento. Se han utilizado las siguientes tecnologías:

- XHTML con componentes de HTML5.
- JavaServerFaces (extensión de PrimeFaces).
- AJAX.
- CSS3 con directivas SASS.

Estas tecnologías aportan dinamismo al front-end y lo hacen más adaptable a nuevas funcionalidades.

6. Se ha incorporado conocimiento experto procedente de los químicos analíticos para buscar los aductos formados por los compuestos en función del modo de ionización utilizado para el análisis de muestras. Los posibles aductos formados para el modo de ionización positivo pueden ser:

   - M+H.
   - M+K.
   - M+Na.
   - M+NH4.
   - M+H-H2O.
   - M+H+HCOONa.
   - M+2H (Aducto con carga doble).

   Y en el modo de ionización negativo:

   - M-H.
   - M+Cl.
   - M+HCOO.
   - M-H-H2O.
   - M-H+HCOONa.

7. Se permite la descarga de resultados en ficheros con formato .xls, formato al que están habituados a trabajar la mayoría de químicos analíticos.

8. Se facilita un análisis de rutas metabolómicas a partir de una hoja de resultados en formato .xls. Esta hoja de resultados debe tener un formato determinado y puede ser de un análisis metabolómico completo o para búsqueda de marcadores biológicos.

9. Se ha incorporado la posibilidad de filtrar resultados en función del origen de datos. Para ello se realiza una búsqueda únicamente en las bases de datos seleccionadas por el usuario.

10. Se permite filtrar los resultados según los elementos químicos que lo forman, con la posibilidad de buscar sobre todos los elementos de la tabla periódica, sobre los elementos CHNOPS, o sobre los elementos CHNOPS+Cl.

11. Se ha realizado una migración de un servidor de aplicaciones obsoleto y con un soporte y desarrollo abandonado por su empresa Oracle como es Glassfish a un servidor Apache TomEE 7.1 recién lanzado por la fundación OpenSource Apache. Se ha configurado el mismo para que se adecúe a las necesidades del proyecto.

El resultado más visible desde un punto de vista externo es la herramienta disponible en la página web Ceu Mass Mediator (http://ceumass.eps.uspceu.es/mediator/), donde se ha integrado toda la funcionalidad desarrollada en este proyecto y está accesibles a través de la interfaz de usuario explicada en el capítulo 5.

6.2. Líneas futuras

Las líneas futuras se dividen en dos secciones diferentes: las líneas futuras en cuanto a las mejoras tecnológicas y las líneas futuras relativas a la identificación de metabolitos.

La primera de las vías tiene una actualización prevista y muy obvia: creación de una API REST para que los usuarios puedan acceder de forma automática a todos los datos y

poder hacer uso de la unificación facilitada mediante este trabajo. Esta línea futura deberá abordarse probablemente con premura, pues hay un equipo de trabajo[70] que ha mostrado interés en integrar nuestra herramienta en un framework desarrollado para poder abordar y compartir de manera abierta estrategias para el análisis de experimentos de metabolómica entre diferentes laboratorios o equipos de trabajo.

Otro desarrollo futuro consistirá en mover la herramienta a un sistema de aprovisionamiento por tiempo como los que se ofrecen actualmente en la nube. Estos sistemas permiten aprovisionar por un tiempo limitado, cuando la petición de usuarios así lo requiera, la máquina con el número de recursos necesarios y facilita las tareas de mantenimiento de software. Teniendo en cuenta que el autor del proyecto es el encargado tanto del desarrollo como del soporte tecnológico, esto podría ocasionar una ganancia de tiempo considerable para las futuras mejoras de la herramienta.

La segunda de las vías debería comenzar con una colaboración con el equipo de Metlin para obtener acceso a sus datos, que se integrarían en nuestra herramienta y proporcionarían una información más completa. Actualmente hay 272.000 metabolitos aproximadamente en esta base de datos, un número mayor que los 166.719 con los que cuenta actualmente nuestra herramienta. Esto es debido a que se están descubriendo continuamente nuevos metabolitos que son añadidos a las diferentes fuentes. Metlin en el 2012 tenía aproximadamente 70.000 compuestos y HMDB entre 3.000 y 4.000 metabolitos. Esto es una muestra de la progresión que está siguiendo esta línea de investigación y la comprobación de que nos encontramos ante un área emergente. En cuanto a la inclusión de estrategias para ofrecer soporte a la identificación de metabolitos, éstas serían algunas de las líneas a explorar:

- Una de las estrategias más prometedoras para dar soporte en el proceso de identificación de metabolitos puede ser el análisis de muestras de composición química conocidas para estudiar los distintos fragmentos, aductos y multímeros generados como consecuencia del protocolo analítico empleado[27]. Una vez esta información es conocida, pueden emplearse herramientas estadísticas capaces de encontrar relaciones entre las

múltiples señales generadas por un mismo metabolito (la generada por el propio metabolito, más las de las distintas alteraciones que pueda sufrir). Estos experimentos también permitirán generar una base de datos conteniendo una lista de posibles reacciones bioquímicas que pueden sufrir los metabolitos como consecuencia del protocolo analítico. Dicha base de datos será empleada por un software que servirá de soporte en la identificación de los distintos metabolitos.

- Otra estrategia para mejorar la identificación de metabolitos es realizar dicha identificación en base a más información que la masa del compuesto. La pieza de información más obvia es el tiempo de retención; esto es, el tiempo que el metabolito tarda en eluir de la fase sólida en el proceso de separación. El tiempo de retención puede proporcionar información muy útil para identificar al metabolito. Lamentablemente, este tiempo depende considerablemente del procedimiento analítico y del montaje experimental[23], lo que hace que sea poco reproducible entre laboratorios, e incluso entre diferentes experimentros dentro del mismo laboratorio, y, por tanto, no se suele encontrar este dato en las bases de datos públicas. Este problema puede mitigarse realizando experimentos con estándares conocidos y construyendo una base de datos interna que contenga información relativa al tiempo de retención. Otra posible estrategia para resolver la carencia de tiempos de retención en bases de datos públicas es emplear simulaciones computacionales para calcular dichos tiempos[31].

- La simulación computacional también puede emplearse para mejorar la identificación de metabolitos a través de la simulación *in silico* de las reacciones químicas que pueden suceder sobre los metabolitos originales como consecuencia del proceso analítico[42]. Estas reacciones dan lugar a compuestos químicos derivados de los metabolitos originales, siendo estos compuestos químicos derivados los que ayudan finalmente identificar los compuestos a partir de los datos generados en el espectrómetro de masas[22]. Una herramienta software capaz de identificar los metabolitos originales a partir de los compuestos químicos derivados que puedan haberse generado como resultado del proceso

analítico sería de gran utilidad[46].

- Cuando todas las estrategias de identificación de los metabolitos fallen, es deseable contar con estrategias para el aislamiento de compuestos no identificados, de acuerdo a las propiedades que se puedan deducir de su comportamiento en las distintas técnicas de separación. Es deseable que dichas estrategias sean pasos explícitos de la técnica analítica y estén soportadas por herramientas de software que reduzcan la carga de trabajo del químico analítico. En concreto, resulta interesante la posibilidad de proporcionar soporte a la búsqueda de compuestos químicos con masas compatibles con los metabolitos identificados en el análisis en PubChem[39], una base de datos que contiene todo tipo de compuestos químicos, no sólo metabolitos. Esto simplificaría y aceleraría el proceso de la identificación de nuevos metabolitos[30].

- En la literatura podemos encontrar autores que abordan el problema de la identificación de metabolitos empleando espectrometría de masas en tándem[6,42,47]. No obstante, los softwares existentes para esta identificación a partir de fragmentos están actualmente en desarrollo y ése es otro de los objetivos a abordar como línea futura de este proyecto.

Bibliografía

[1] A. Alonso, A. Julià, S. Marsal, A. Beltran, M. Vinaixa, X. Correig, M. Díaz, and L. Ibañez. Astream: An r package for annotating lc/ms metabolomic data. *Bioinformatics*, 27(9):1339–1340, / 05 / 01 / 2011. ID: edselc.2-52.0-79954462033; M2: 1339; Accession Number: edselc.2-52.0-79954462033; (Bioinformatics, May 2011, 27(9):1339-1340) Publication Type: Academic Journal; Rights: Copyright 2011 Elsevier B.V., All rights reserved.

[2] Balsamiq. Balsamiq, 2016. https://balsamiq.com/ [Online; accedido 15-06-2016].

[3] Bitbucket. Bitbucket, 2016. https://bitbucket.org/ [Online; accedido 15-06-2016].

[4] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses, and W. B. Dunn. Automated workflows for accurate mass-based putative metabolite identification in lc/ms-derived metabolomic datasets. *Bioinformatics (Oxford, England)*, 27(8):1108–1112, Apr 15 2011.

[5] Marie Brown, Warwick B. Dunn, P. Dobson, Y. Patel, CL Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, and A. Tseng. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7):1322–1332, 2009.

[6] Mingshu Cao, Karl Fraser, and Susanne Rasmussen. Computational analyses of spectral trees from electrospray multi-stage mass spectrometry to aid metabolite identification. *Metabolites*, 3(4):1036–1050, 2013.

[7] Dalgliesh C.E., Horning M.G. Horning E.C., and Yarger K. Knox K.L. A gas-liquid-chromatographic procedure for separating a wide range of metabolites occurring in urine or tissue extracts. *Biochemical Journal*, 101:792, 1966.

[8] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data, 2011. ID: 000285831700109.

[9] A. (. 1.). Chokkathukalam, Jankevics, ¡ A. (1, 2), F. (. 1.). Achcar, R. (1 Breitling, 2, 5), Creek, D.J. (3, 4), and M. P. (. 4.). Barrett. Mzmatch-iso: An r tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics*, 29(2):281–283, / 01 / 15 / 2013. ID: edselc.2-52.0-84872544333; M2: 281; Accession Number: edselc.2-52.0-84872544333; (Bioinformatics, 15 January 2013, 29(2):281-283) Publication Type: Academic Journal; Rights: Copyright 2013 Elsevier B.V., All rights reserved.

[10] R. Daly, S. Rogers, J. Wandy, A. Jankevics, KEV Burgess, and R. Breitling. Metassign: probabilistic annotation of metabolites from lc-ms data using a bayesian clustering approach, 2014. ID: 000343082900010.

[11] Kevin Davies. The human metabolome project. *Bio-IT World*, 6(3):6, 04 2007. 24833233.

[12] Andros Corral Paya. Universidad de Valencia. Facultad de Farmacia. Departamento de Quimica Analitica. Fundamentos y funciones de la espectrometría de masas, 2006. http://mural.uv.es/caloan/ [Online; accedido 15-06-2016].

[13] Warwick B. Dunn, Alexander Erban, Ralf JM Weber, Darren J. Creek, Marie Brown, Rainer Breitling, Thomas Hankemeier, Royston Goodacre, Steffen Neumann, and Joachim Kopka. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(1):44–66, 2013.

[14] Git. Git, 2016. https://git-scm.com/ [Online; accedido 15-06-2016].

[15] Diana Gonzalez-Pena, Danuta Dudzik, Clara Colina-Coca, Begona Ancos, Antonia Garcia, Coral Barbas, and Concepcion Sanchez-Moreno. Multiplatform metabolomic finger-

printing as a tool for understanding hypercholesterolemia in wistar rats. *European journal of nutrition*, (3):997, 2016. ID: Accession Number: edsgcl.448244165; Item Citation: European Journal of Nutrition. April 2016, Vol. 55 Issue 3, p997, 14 p.; Accession Number: edsgcl.448244165; Publication Type: Academic Journal; Source: European Journal of Nutrition; Language: English; Publication Date: 20160401; Rights: Copyright 2016 Gale, Cengage Learning. All rights reserved., COPYRIGHT 2016 Springer; Imprint: Springer.

[16] Jennifer Griffiths. A brief history of mass spectrometry. *Analytical Chemistry*, 80(15):5678–5683, 8 2008.

[17] herausgegeben von S. Dagley and Donald E. Nicholson. An introduction to metabolic pathways. *Seiten, zahlreiche Abb*, 15(4):473, 1970.

[18] HTML5UP. Html5up, 2016. https://html5up.net/ [Online; accedido 15-06-2016].

[19] Caroline H. Johnson and Frank J. Gonzalez. Challenges and opportunities of metabolomics. *Journal of cellular physiology*, 227(8):2975–2981, 2012.

[20] Lakeworks. The scrum project management method. https://upload.wikimedia.org/wikipedia/commons/5/58/Scrum_process.svg [Online; accedido 22-08-2016].

[21] Siuzdak G. et al. Lerner R.A. Cerebrodiene: a brain lipid isolated from sleep-deprived cats. *Proc Natl Acad Sci*, 91(20):9505–9508, 1994.

[22] Liang Li, Ronghong Li, Jianjun Zhou, Azeret Zuniga, Avalyn E. Stanislaus, Yiman Wu, Tao Huan, Jiamin Zheng, Yi Shi, and David S. Wishart. Mycompoundid: using an evidence-based metabolome library for metabolite identification. *Analytical Chemistry*, 85(6):3401–3408, 2013.

[23] John C. Lindon, Jeremy K. Nicholson, and Elaine Holmes. *The handbook of metabonomics and metabolomics.* Amsterdam etc.] : Elsevier, 2007, 2007.

[24] LipidMaps. Lipidmaps, 2016. http://www.lipidmaps.org/ [Online; accedido 09-06-2016].

[25] Michael Lämmerhofer and Wolfram Weckwerth. *Metabolomics in practice: successful strategies to generate and analyze metabolic data.* John Wiley & Sons, 2013.

[26] A. Lommen, H. J. Kools, H. Gowda, J. Ivanisevic, C. H. Johnson, M. E. Kurczy, H. P. Benton, D. Rinehart, T. Nguyen, J. Ray, J. Kuehl, B. Arevalo, P. D. Westenskow, J. H. Wang, A. P. Arkin, A. M. Deutschbauer, G. J. Patti, and G. Siuzdak. Metalign 3.0: Performance enhancement by efficient use of advances in computer hardware, / 08 / 01 / 2014. TY: JOUR; ID: edselc.2-52.0-84864071589; J1: Metabolomics; M2: 719; Accession Number: edselc.2-52.0-84864071589; (Metabolomics, August 2012, 8(4):719-726) Publication Type: Academic Journal; Rights: Copyright 2012 Elsevier B.V., All rights reserved.; ID: 000339227400031.

[27] Ke-Shiuan Lynn, Mei-Ling Cheng, Yet-Ran Chen, Chin Hsu, Ann Chen, T. Mamie Lih, Hui-Yin Chang, Ching jang Huang, Ming-Shi Shiao, and Wen-Harn Pan. Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Analytical Chemistry*, 87(4):2143–2151, 2015.

[28] Dole M., Hines R.L. Mack L.L., Ferguson L.D. Mobley R.C., and Alice M.B. Molecular beams of macroions. *Journal of Chemical Physics*, 49(5):2240, 1968.

[29] Margaret Anne Madigan, Kenneth Blum, and Debmalya Barh. *Omics : Biomedical Perspectives and Applications.* CRC Press, Boca Raton, 2012. ID: Accession Number: 473322.

[30] Lochana C. Menikarachchi, Shannon Cawley, Dennis W. Hill, L. Mark Hall, Lowell Hall, Steven Lai, Janine Wilder, and David F. Grant. Molfind: a software package enabling

hplc/ms-based identification of unknown chemical structures. *Analytical Chemistry*, 84(21):9388–9394, 2012.

[31] Lochana C. Menikarachchi, Mai A. Hamdalla, Dennis W. Hill, and David F. Grant. Chemical structure identification in metabolomics: computational modeling of experimental features. *Computational and structural biotechnology journal*, 5(6):1–7, 2013.

[32] Metlin. Metlin, 2016. https://metlin.scripps.edu/index.php [Online; accedido 09-06-2016].

[33] Eric Milgram and Anders Nordstrom. Asms metabolomics workshop: Current topics in metabolomics, 2009.

[34] Field F.H. Munson M.S.B. Chemical ionization mass spectrometry i. general introduction. *J. Am. Chem. Soc.*, 88(12):2621–2630, 1966.

[35] MzMatch. Peakml schema, 2016. http://mzmatch.sourceforge.net/peakml_specification.pdf [Online; accedido 09-06-2016].

[36] Alicia Navarrete, Emily G. Armitage, Monica Musteanu, Antonia García, Annalaura Mastrangelo, Renata Bujak, Pedro P. López-Casas, Manuel Hidalgo, and Coral Barbas. Metabolomic evaluation of mitomycin c and rapamycin in a personalized treatment of pancreatic cancer. *Pharmacology Research & Perspectives*, 2(6):n/a–n/a, 12 2014.

[37] Fiehn O. Metabolomics-the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.

[38] Yanes O. La metabolómica: un deja vu por la historia de la bioquimica. *SEBBM*, 186:4–7, 2015.

[39] National Institute of General Medical Sciences. Pubchemical, 2016. https://pubchem.ncbi.nlm.nih.gov/ [Online; accedido 09-06-2016].

[40] Kyoto Encyclopedia of Genes and Genomes. Kegg, 2016. http://www.genome.jp/kegg [Online; accedido 09-06-2016].

[41] Canadian Institutes of Health Research. Hmdb, 2016. http://www.hmdb.ca/ [Online; accedido 09-06-2016].

[42] Julio E. Peironcely, Miguel Rojas-Cherto, Albert Tas, Rob Vreeken, Theo Reijmers, Leon Coulier, and Thomas Hankemeier. Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics. *Analytical Chemistry*, 85(7):3576–3583, 2013.

[43] Leonid Poretsky. *Principles of Diabetes Mellitus*, volume 2nd ed. Springer, New York, 2010. ID: Accession Number: 313048.

[44] P.ravisankar. Mass spectrometry (mass-spec), 2013. http://www.slideshare.net/banuman35/mass-spectrometrymassspec2013-pravisankar página 30 [Online; accedido 22-08-2016].

[45] PrimeFaces. Primefaces, 2016. http://primefaces.org/ [Online; accedido 15-06-2016].

[46] S. Rogers, R. A. Scheltema, M. Girolami, and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics (Oxford, England)*, 25(4):512–518, Feb 15 2009. LR: 20160322; GR: BB/G006997/1/Biotechnology and Biological Sciences Research Council/United Kingdom; GR: G0401466/Medical Research Council/United Kingdom; JID: 9808944; 0 (Proteome); 2008/12/18 [aheadofprint]; ppublish.

[47] Miquel Rojas-Cherto, Julio E. Peironcely, Piotr T. Kasper, Justin JJ van der Hooft, Ric CH de Vos, Rob Vreeken, Thomas Hankemeier, and Theo Reijmers. Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Analytical Chemistry*, 84(13):5524–5534, 2012.

[48] D. Rojo, C. Barbas, and F. J. Ruperez. Lc-ms metabolomics of polar compounds, 2012. ID: 000306161000017.

[49] Nathan Salomonis, Kristina Hanspers, Alexander C. Zambon, Karen Vranizan, Steven C. Lawlor, Kam D. Dahlquist, Scott W. Doniger, Josh Stuart, Bruce R. Conklin, and Alexander R. Pico. Genmapp 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217–12, 01 2007.

[50] NG De Santo, M. Cirillo, C. Bisaccia, A. Mezzogiorno, R. Pisot, and G. Ongaro. Twenty-six renal aphorisms of Santorio Santorio (1561-1636), 2013. ID: 000329765900006.

[51] Augustin Scalbert, Lorraine Brennan, Oliver Fiehn, Thomas Hankemeier, Bruce S. Kristal, Ben van Ommen, Estelle Pujos-Guillot, Elwin Verheij, David Wishart, and Suzan Wopereis. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5(4):435–458, 2009.

[52] Richard A. Scheltema, Saskia Decuypere, Jean-Claude Dujardin, David G. Watson, Ritsert C. Jansen, and Rainer Breitling. Simple data-reduction method for high-resolution lc-ms data in metabolomics. *Bioanalysis*, 1(9):1551–1557, 2009.

[53] Richard A. Scheltema, Andris Jankevics, Ritsert C. Jansen, Morris A. Swertz, and Rainer Breitling. Peakml/mzmatch: a file format, java library, r library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry*, 7(2786), 2011. ID: Accession Number: edsgcl.254678023; Item Citation: Analytical Chemistry. April 1, 2011, Vol. 83 Issue 7, p2786, 8 p.; Accession Number: edsgcl.254678023; Publication Type: Academic Journal; Source: Analytical Chemistry; Language: English; Publication Date: 20110401; Rights: Copyright 2011 Gale, Cengage Learning. All rights reserved., COPYRIGHT 2011 American Chemical Society; Imprint: American Chemical Society.

[54] K. S. Sharma. Mass spectrometry—the early years. *International Journal of Mass*

Spectrometry, 349-350(100):3–8, 2013. ID: S1387380613002145; Accession Number: S1387380613002145; Author: Sharma, K.S.; Affiliation: Department of Physics and Astronomy, University of Manitoba, Winnipeg, Canada R3T 2N2; Number of Pages: 6; Language: English;.

[55] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. Metlin - a metabolite mass spectral database, 2005. ID: 000233812400016.

[56] R. Smith, A. D. Mathis, D. Ventura, and J. T. Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC bioinformatics*, 15 Suppl 7:S9, / 01 / 01 / 2014. ID: edselc.2-52.0-84907424497; M2: S9; Accession Number: edselc.2-52.0-84907424497; (BMC bioinformatics, 2014, 15 Suppl 7:S9) Publication Type: Academic Journal; Rights: Copyright 2014 Medline is the source for the citation and abstract of this record.

[57] F. Sobott, J. L. Benesch, E. Vierling, and C. V. Robinson. Subunit exchange of multimeric protein complexes. real-time monitoring of subunit exchange between small heat shock proteins by using electrospray mass spectrometry. *The Journal of biological chemistry*, 277(41):38921–38929, Oct 11 2002. LR: 20071114; GR: GM-42762/GM/NIGMS NIH HHS/United States; JID: 2985121R; 0 (HSP16.9 protein, Triticum aestivum); 0 (Heat-Shock Proteins); 0 (Macromolecular Substances); 0 (Plant Proteins); 0 (Protein Subunits); 2002/07/23 [aheadofprint]; publish.

[58] American Chemical Society. Cas, 2016. https://www.cas.org [Online; accedido 09-06-2016].

[59] Symyx Solutions. Mol files manual, 2016. http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/MDL_SDF.pdf [Online; accedido 09-06-2016].

[60] K. Strimbu and J. A. Tavel. What are biomarkers?, 2010. ID: 000295509900002.

[61] R. (. 1.). Tautenhahn, D. (. 1.). Rinehart, G. (. 1.). Siuzdak, and G. J. (. 2.). Patti. Xcms online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11):5035–5039, / 06 / 05 / 2012. ID: edselc.2-52.0-84861915827; M2: 5035; Accession Number: edselc.2-52.0-84861915827; (Analytical Chemistry, 5 June 2012, 84(11):5035-5039) Publication Type: Academic Journal; Rights: Copyright 2013 Elsevier B.V., All rights reserved.

[62] InChI Trust. Inchi software, 2016. http://www.inchi-trust.org/downloads/ [Online; accedido 09-06-2016].

[63] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences®*, 28:31–36, / 01 / 01 / 1988. ID: edselc.2-52.0-0023965741; M2: 31; Accession Number: edselc.2-52.0-0023965741; (Journal of Chemical Information and Computer Sciences®, 1988, 28:31-36) Publication Type: Academic Journal; Rights: Copyright 2006 Elsevier B.V., All rights reserved.

[64] R.J. et al Williams. Individual metabolic patterns and human disease: An exploratory study utilizing predominantly paper chromatographic methods. *U. Texas Publication*, 5109(204), 1951.

[65] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. F. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. G. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. Hmdb 3.0-the human metabolome database in 2013, 2013. ID: 000312893300113.

[66] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe,

P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. Hmdb: the human metabolome database. *Nucleic acids research*, 35(Database issue):D521–6, Jan 2007. LR: 20140907; JID: 0411011; OID: NLM: PMC1899095; ppublish.

[67] David S. Wishart. Advances in metabolite identification. *Bioanalysis*, 3(15):1769–1782, 2011.

[68] David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De Souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythe. Hmdb: a knowledgebase for the human metabolome. *Nucleic acids research*, 2009. ID: Accession Number: edsagr.US201301569821; Item Citation: Nucleic acids research. 2009 Jan., v. 37, no. suppl_1; Accession Number: edsagr.US201301569821; Publication Type: Periodical; Source: Nucleic acids research; Language: English; Format: text; Publication Date: 20090101.

[69] DW Woolley and AGC White. Production of thiamine deficiency disease by the feeding of a pyridine analogue of thiamine. *The Journal of Biological Chemistry*, July(149):285–289, 1943.

[70] WorkFlow4Metabolomics. Workflow4metabolomics, 2016. http://workflow4metabolomics.org/ [Online; accedido 13-08-2016].

[71] J. (. 1.). Xia, Wishart D.S. (1, 2, and 3). Metpa: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, 26(18):2342–2344, / 07 / 13 / 2010. ID: edselc.2-52.0-77956550520; M2: 2342; Accession Number: edselc.2-52.0-77956550520; (Bioinformatics, 13 July 2010, 26(18):2342-2344) Publication Type: Academic Journal; Rights: Copyright 2010 Elsevier B.V., All rights reserved.

Apéndice A

Descomposición de molécula mediante espectrometría en tándem

La figura A.1 muestra el espectro compuesto del metabolito, cuya molécula inicial se puede observar en la figura A.2. Este elemento de la figura A.2 es el 1-Methoxy-1-pentyloxyethane con CAS: 73142-32-2. En las figuras que van desde la A.3 a la A.10 se puede observar los compuestos con mayor intensidad generados al aplicar espectrometría en tandem sobre el metabolito inicial.

Esta técnica sirve para identificar elementos cuya información previa (masa experimental y tiempo de retención) no es suficiente para averiguar el metabolito del que se trata. En ese caso hay que acudir a otro tipo de técnicas y MS/MS es una de ellas. Se observa que la molécula rompe determinados enlaces en cantidades conocidas, y eso es una información muy relevante a la hora de la identificación.

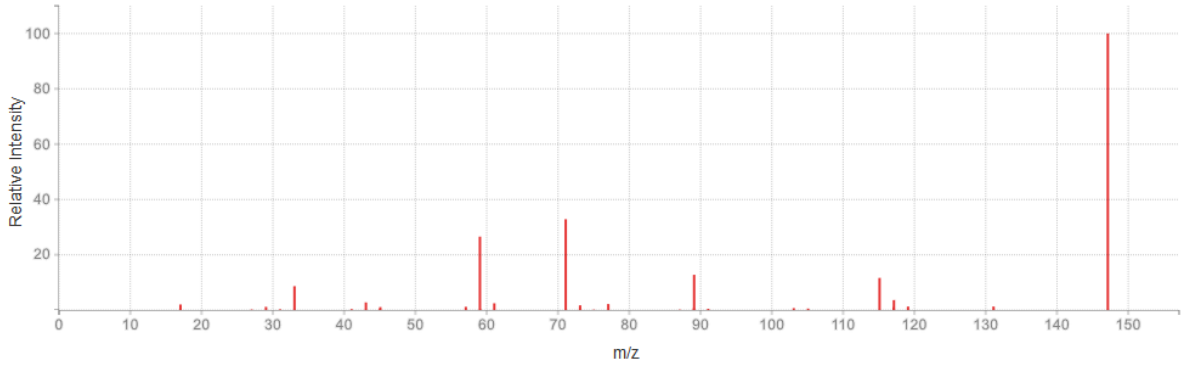


Figura A.1: *Espectro compuesto del metabolito 1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2*[41]

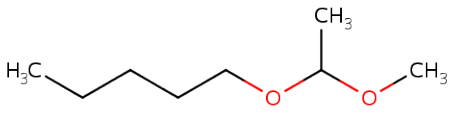


Figura A.2: *Molécula inicial (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2) sobre la que se aplica espectrometría en tándem*[41]

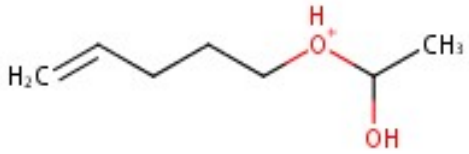


Figura A.3: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

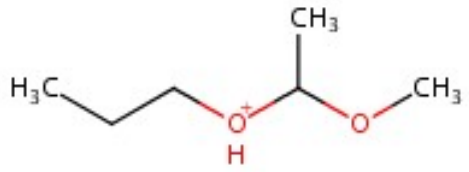


Figura A.4: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

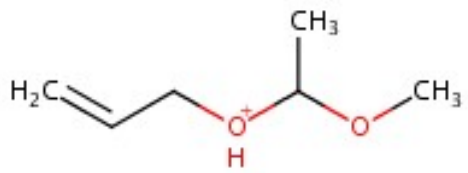


Figura A.5: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

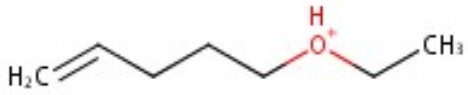


Figura A.6: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

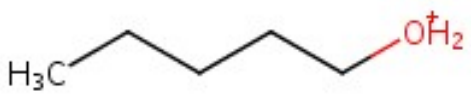


Figura A.7: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

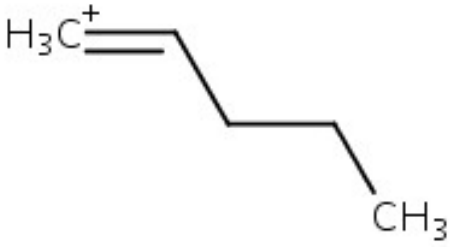


Figura A.8: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

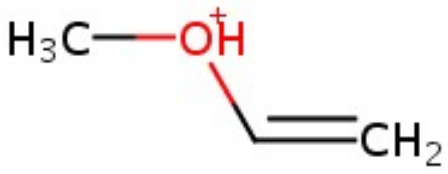


Figura A.9: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

Figura A.10: *Molécula obtenida tras aplicar espectrometría en tándem al elemento (1-Methoxy-1-pentyloxyethane, CAS: 73142-32-2)*[41]

Apéndice B

Script de generación de identificadores InChI

A continuación se muestra el código fuente del script que sea empleado para la generación de los identificadores InChi.

```
#!/bin/bash
set -x
#download files from LipidMaps and HMDB and extract the files
   which the project is working with.
# WARNING: LOOK IF THE FILE ON THE WEB PAGES IS THE LAST
   AVAILABLE
# AND CHANGE THE DATE OF THE FILE ON LIPIDMAPS IN THE FIRST
   VARIABLE

#Function to Retrieve the mol Files from Kegg
# Function to generate KeggCode
generateKeggCode () {
variable=10
if [ $1 -lt $variable ]; then
eval $2=C0000$1
else
variable=100
if [ $1 -lt $variable ]; then
eval $2=C000$1
else
variable=1000
if [ $1 -lt $variable ]; then
eval $2=C00$1
else
```

```
variable=10000
if [ $1 -lt $variable ]; then
eval $2=C0$1
else
eval $2=C$1
fi
fi
fi
fi
}
#Function do DownloadMolFiles
downloadMolFiles () {
for i in `seq 1 $1`
do
fileName=''
generateKeggCode $i fileName
#    echo $fileName
wget --tries=50 http://www.genome.jp/dbget-bin/www_bget?-f+m+
   compound+$fileName -P ./ceuMassUpdating/resources/kegg/
   molfiles
mv ./ceuMassUpdating/resources/kegg/molfiles/www_bget?-f+m+
   compound+$fileName ./ceuMassUpdating/resources/kegg/molfiles
   /$fileName.mol
done
}

generateInChIFromMol()
{
/home/alberto/PHD/mediator/inchi/INCHI-1-BIN/linux/64bit/inchi
   -1 $1 $1.txt -key
}

generateKeggInChIs()
{
for file in ./ceuMassUpdating/resources/kegg/molfiles/*mol
do
generateInChIFromMol $file
done
}

#Variables
dateExecution=$(date +"%y%m%d")
dateLMFile=LMSDFDownload28Jun15
```

```
nameLMFileToDelete=FinalAll.sdf
numberDBKegg=6
numberDBHMDB=18
numberDBPubChem=22
sourceSeparation=src
numberCompoundsKegg=21200
nameFileKeggHMDB=
   $sourceSeparation$numberDBKegg$sourceSeparation$numberDBHMDB
   .txt.gz
nameFileKeggPubChem=
   $sourceSeparation$numberDBKegg$sourceSeparation$numberDBPubChem
   .txt.gz
nameFileHMDBPubChem=
   $sourceSeparation$numberDBHMDB$sourceSeparation$numberDBPubChem
   .txt.gz
#To test the script it is created a directory test which is
   necessary to change for the final version
test=test
#Commands
#download molFiles from Kegg
cp -R ./ceuMassUpdating/ ./ceuMassUpdating$dateExecution
rm -Rf ./ceuMassUpdating/resources/kegg/molfiles/*
downloadMolFiles $numberCompoundsKegg
generateKeggInChIs

#download files from LipidMaps
rm -Rf ./ceuMassUpdating/resources/$test/lipidMaps/*
wget --tries=50 www.lipidmaps.org/resources/downloads/
   $dateLMFile.tar.gz -P ./ceuMassUpdating/resources/$test/
   lipidMaps
gunzip ./ceuMassUpdating/resources/$test/lipidMaps/$dateLMFile.
   tar.gz
tar -xvf ./ceuMassUpdating/resources/$test/lipidMaps/
   $dateLMFile.tar -C ./ceuMassUpdating/resources/$test/
   lipidMaps/
rm ./ceuMassUpdating/resources/$test/lipidMaps/$dateLMFile.tar
   ./ceuMassUpdating/resources/$test/lipidMaps/$dateLMFile/
   $dateLMFile$nameLMFileToDelete
mv ./ceuMassUpdating/resources/$test/lipidMaps/$dateLMFile/* ./
   ceuMassUpdating/resources/$test/lipidMaps/
rmdir ./ceuMassUpdating/resources/$test/lipidMaps/$dateLMFile
#download files from HMDB
for i in ls ./ceuMassUpdating/resources/$test/HMDB/*; do rm -v
```

```
   $i -f; done
wget --tries=50 www.hmdb.ca/system/downloads/current/
   hmdb_metabolites.zip -P ./ceuMassUpdating/resources/$test/
   HMDB
unzip ./ceuMassUpdating/resources/$test/HMDB/hmdb_metabolites.
   zip -d ./ceuMassUpdating/resources/$test/HMDB/
rm ./ceuMassUpdating/resources/$test/HMDB/hmdb_metabolites.zip
   ./ceuMassUpdating/resources/$test/HMDB/hmdb_metabolites.xml
#download files from UniChem
wget --tries=50 ftp://ftp.ebi.ac.uk/pub/databases/chembl/
   UniChem/data/wholeSourceMapping/src_id$numberDBKegg/
   $nameFileKeggHMDB -P ./ceuMassUpdating/resources/$test/
   unichem
gunzip ./ceuMassUpdating/resources/$test/unichem/
   $nameFileKeggHMDB
wget --tries=50 ftp://ftp.ebi.ac.uk/pub/databases/chembl/
   UniChem/data/wholeSourceMapping/src_id$numberDBKegg/
   $nameFileKeggPubChem -P ./ceuMassUpdating/resources/$test/
   unichem
gunzip ./ceuMassUpdating/resources/$test/unichem/
   $nameFileKeggPubChem
wget --tries=50 ftp://ftp.ebi.ac.uk/pub/databases/chembl/
   UniChem/data/wholeSourceMapping/src_id$numberDBHMDB/
   $nameFileHMDBPubChem -P ./ceuMassUpdating/resources/$test/
   unichem
gunzip ./ceuMassUpdating/resources/$test/unichem/
   $nameFileHMDBPubChem

#remove files from ChemIdPlus
for i in ls ./ceuMassUpdating/resources/$test/cas/*; do rm -v
   $i -f; done
#The download of the files is done in the java program
#remove the previous compounds from kegg
for i in ls ./ceuMassUpdating/resources/$test/kegg/*; do rm -v
   $i -f; done
#clear the log
/dev/null > ./ceuMassUpdating/log/kegg_duplicates.txt
#/dev/null > ./ceuMassUpdating/log/*
#execute the java project to download kegg compounds and
   populate databases
java -jar "/home/alberto/PHD/mediator/DBupdater/ceuMassUpdating
   /dist/ceuMassUpdating.jar"
```